日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年１０月２８日

出願番号
Application Number: 特願２００２－３１２３３１

[ST.10/C]: [ＪＰ２００２－３１２３３１]

出願人
Applicant(s): インターナショナル・ビジネス・マシーンズ・コーポレーション

２００３年　５月２０日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０３７００２

【書類名】　特許願

【整理番号】　JP9020152

【あて先】　特許庁長官殿

【国際特許分類】　G06F 17/30

G06F 3/14

G06F 13/00

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

【氏名】　▲高▼木　啓伸

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

【氏名】　浅川　千恵子

【特許出願人】

【識別番号】　390009531

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】　100086243

【弁理士】

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【弁理士】

【氏名又は名称】　市位　嘉宏

【代理人】

【識別番号】　100108501

【弁理士】

【氏名又は名称】　上野　剛史

【復代理人】

【識別番号】　100085408

【弁理士】

【氏名又は名称】　山崎　隆

【手数料の表示】

【予納台帳番号】　117560

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9706050

【包括委任状番号】　9704733

【包括委任状番号】　0207860

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　構造化・階層化コンテンツ用処理装置、構造化・階層化コンテンツ用処理方法、及びプログラム

【特許請求の範囲】

【請求項１】　ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う構造化・階層化コンテンツ用処理装置であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定手段、

前記ターゲット・コンテンツに対する過去の複数個の構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出手段、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成手段、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類手段、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成手段、

を有していることを特徴とする構造化・階層化コンテンツ用処理装置。

【請求項２】　前記所定の処理とは、該構造化・階層化コンテンツのコンテンツ部分への関連情報の関連付けであることを特徴とする請求項１記載の構造化・階層化コンテンツ用処理装置。

【請求項３】　前記関連情報はアノテーションを含むことを特徴とする請求項２記載の構造化・階層化コンテンツ用処理装置。

【請求項４】　前記所定の処理とは、構造化・階層化コンテンツのコンテンツ部分を他の構造化・階層化コンテンツに利用するために該構造化・階層化コンテンツの該コンテンツ部分をコピーする処理であることを特徴とする請求項１記載の構造化・階層化コンテンツ用処理装置。

【請求項５】　構造化・階層化コンテンツとはウェブ・コンテンツであることを特徴とする請求項１記載の構造化・階層化コンテンツ用処理装置。

【請求項６】　ターゲット・サブツリーのノードを、定常ノード、更新ノード及び付加ノードに分類する前記分類手段を有していることを特徴とする請求項１記載の構造化・階層化コンテンツ用処理装置。

【請求項７】　検出する前記出現態様として、（Ｎ１）被検出ノードがターゲット・コンテンツ部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に同一となって出現態様、及び（Ｎ２）被検出ノードがターゲット・コンテンツ部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に異なっている出現態様を含む前記出現態様検出手段、及び

統計情報により（Ｎ１）の出現態様による出現頻度が第１の閾値以上であると判明したノードは定常ノードに分類し、統計情報により（Ｎ２）の出現態様による出現頻度が第２の閾値以上であると判明したノードは更新ノードに分類し、定常ノード及び更新ノード以外のノードは付加ノードに分類する前記分類手段、

を有していることを特徴とする請求項６記載の構造化・階層化コンテンツ用処理装置。

【請求項８】　前記マッチング・パターン生成手段は、

定常ノード、更新ノード及び付加ノードの分類に基づいてターゲット・サブツリーにおける繰り返し部分を検出する繰り返し部分検出手段、及び

該繰り返し部分の存在情報を含む前記マッチング・パターンを生成する繰り返し情報付きマッチング・パターン生成手段、

を有していることを特徴とする請求項６記載の構造化・階層化コンテンツ用処理装置。

【請求項９】　前記分類手段は、

イメージに係るノードについて、該ノードが空白領域を確保するためのスペー

サ用イメージに係るノードであるか否かを検出するスペーサ用イメージ検出手段、

イメージに係るノードについて、該ノードが繰り返して同一サイズで複数個使用されるビュレット・イメージに係るノードであるか否かを検出するビュレット・イメージ検出手段、

スペーサ用イメージに係るノードは付加ノードと分類する第１の分類付け手段、

ビュレット・イメージに係るノード同士は、その表示内容が異なっても定常ノード、更新ノード又は付加ノードの同一分類に割り当てる第２の分類付け手段、

を有していることを特徴とする請求項８記載の構造化・階層化コンテンツ用処理装置。

【請求項１０】　ターゲット・コンテンツに対する過去の構造化・階層化コンテンツが存在しない場合には、過去の各構造化・階層化コンテンツの代わりに該ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択しターゲット・コンテンツに係るターゲット・サブツリーと各隣接構造化・階層化コンテンツに係るツリーと対照する前記対照手段、

を有していることを特徴とする請求項１記載の構造化・階層化コンテンツ用処理装置。

【請求項１１】　ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う構造化・階層化コンテンツ用処理装置であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定手段、

前記ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと各隣接構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各

ノードの出現態様を検出する出現態様検出手段、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成手段、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類手段、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成手段、

を有していることを特徴とする構造化・階層化コンテンツ用処理装置。

【請求項１２】　ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う構造化・階層化コンテンツ用処理方法であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定ステップ、

前記ターゲット・コンテンツに対する過去の複数個の構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出ステップ、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成ステップ、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類ステップ、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成ステップ、

を有していることを特徴とする構造化・階層化コンテンツ用処理方法。

【請求項１３】　前記所定の処理とは、該構造化・階層化コンテンツのコンテンツ部分への関連情報の関連付けであることを特徴とする請求項１２記載の構造化・階層化コンテンツ用処理方法。

【請求項１４】　前記関連情報はアノテーションを含むことを特徴とする請求項１３記載の構造化・階層化コンテンツ用処理方法。

【請求項１５】　前記所定の処理とは、構造化・階層化コンテンツのコンテンツ部分を他の構造化・階層化コンテンツに利用するために該構造化・階層化コンテンツの該コンテンツ部分をコピーする処理であることを特徴とする請求項１２記載の構造化・階層化コンテンツ用処理方法。

【請求項１６】　構造化・階層化コンテンツとはウェブ・コンテンツであることを特徴とする請求項１２記載の構造化・階層化コンテンツ用処理方法。

【請求項１７】　ターゲット・サブツリーのノードを、定常ノード、更新ノード及び付加ノードに分類する前記分類ステップを有していることを特徴とする請求項１２記載の構造化・階層化コンテンツ用処理方法。

【請求項１８】　検出する前記出現態様として、（Ｎ１）被検出ノードがターゲット・コンテンツ部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に同一となって出現態様、及び（Ｎ２）被検出ノードがターゲット・コンテンツ部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に異なっている出現態様を含む前記出現態様検出ステップ、及び

統計情報により（Ｎ１）の出現態様による出現頻度が第１の閾値以上であると判明したノードは定常ノードに分類し、統計情報により（Ｎ２）の出現態様による出現頻度が第２の閾値以上であると判明したノードは更新ノードに分類し、定常ノード及び更新ノード以外のノードは付加ノードに分類する前記分類ステップ

を有していることを特徴とする請求項１７記載の構造化・階層化コンテンツ用処理方法。

【請求項１９】　前記マッチング・パターン生成ステップは、

定常ノード、更新ノード及び付加ノードの分類に基づいてターゲット・サブツ

リーにおける繰り返し部分を検出する繰り返し部分検出ステップ、及び

該繰り返し部分の存在情報を含む前記マッチング・パターンを生成する繰り返し情報付きマッチング・パターン生成ステップ、

を有していることを特徴とする請求項１７記載の構造化・階層化コンテンツ用処理方法。

【請求項２０】　前記分類ステップは、

イメージに係るノードについて、該ノードが空白領域を確保するためのスペーサ用イメージに係るノードであるか否かを検出するスペーサ用イメージ検出ステップ、

イメージに係るノードについて、該ノードが繰り返して同一サイズで複数個使用されるビュレット・イメージに係るノードであるか否かを検出するビュレット・イメージ検出ステップ、

スペーサ用イメージに係るノードは付加ノードと分類する第１の分類付けステップ、

ビュレット・イメージに係るノード同士は、その表示内容が異なっても定常ノード、更新ノード又は付加ノードの同一分類に割り当てる第２の分類付けステップ、

を有していることを特徴とする請求項１９記載の構造化・階層化コンテンツ用処理方法。

【請求項２１】　ターゲット・コンテンツに対する過去の構造化・階層化コンテンツが存在しない場合には、過去の各構造化・階層化コンテンツの代わりに該ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択しターゲット・コンテンツに係るターゲット・サブツリーと各隣接構造化・階層化コンテンツに係るツリーと対照する前記対照ステップ、

を有していることを特徴とする請求項１２記載の構造化・階層化コンテンツ用処理方法。

【請求項２２】　ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行

う構造化・階層化コンテンツ用処理方法であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定ステップ、

前記ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと各隣接構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出ステップ、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成ステップ、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類ステップ、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成ステップ、

を有していることを特徴とする構造化・階層化コンテンツ用処理方法。

【請求項２３】　請求項１２～２２のいずれかに記載の構造化・階層化コンテンツ用処理方法の各ステップをコンピュータに実行させるプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、アノテーションの使い回しやウェブ・コンテンツの切り出し等の処理に適する構造化・階層化コンテンツ用処理装置、構造化・階層化コンテンツ用処理方法、及び構造化・階層化コンテンツ用処理プログラムに係り、詳しくは、アノテーションの使い回しやウェブ・コンテンツの切り出し等の処理対象の構造化・階層化コンテンツを適切に検出できるマッチング・パターンを生成できる生成する構造化・階層化コンテンツ用処理装置、構造化・階層化コンテンツ用処理

方法、及び構造化・階層化コンテンツ用処理プログラムに関するものである。

【０００２】

【従来の技術】

近年、大量に存在するウェブ・ページから重要なコンテンツを含む部分を切り出してパーツ化することにより、高度に再利用する研究が様々な観点から注目されている。なお、本明細書において、「切り出し」とは、当業者において一般的に使用されている意味で使用しており、該切り出しによって切り出し元のウェブ・コンテンツから切り出し部分が削除されることはない。本明細書における「切り出し」とは、厳密に言うと、別のウェブ・ページ等に対象のコンテンツ部分を貼り付けるために、オリジナルのウェブ・コンテンツ等において対象のコンテンツ部分の範囲をコピーすることである。

【０００３】

Web Serviceの分野では、既存のHTMLコンテンツとWeb Serviceの橋渡しをするブリッジング・テクノロジーとしてコンテンツ切り出しが注目されている。例えば、ニュース・サイトの記事をキーワード検索するHTMLフォームを切り出してXML入出力を定義することで、既存のサーバ・システムをそのままにWebサービス化することができる。

【０００４】

また、様々な情報を統合(aggregate)してユーザの要求に合致したポータル・ページを提供する情報ポータル（Information Portal）の分野では、既存ウェブ・ページの部分コンポーネントは重要なコンテンツである。様々なニュース・サイトからトップ・ニュースやヘッドラインの領域を切り出して自由に組み合わせることでコンテンツを飛躍的に拡大させることができる。実際にmySiteOutliner、WebSphere Portal Server等では既存ウェブ・ページの一部をポータル・ページに組み込む仕組みが製品の一部として提供されている。

【０００５】

また、ウェブ・サイトで更新された情報等をRSS(Rich Site Summary)というXML形式で提供することにより、第３者が利用できるようにする規格が広まってきている。現在はRSSは専用のサーバサードプログラム（CGI等）を用意することで

生成されているが、ページ切り出し技術を用いれば、ページ内のヘッドライン・リストをRSSに変換することでダイナミックかつ、即時性の高いRSSを提供可能である。

【０００６】

さらにトランスコーディングの分野では、ページ内の重要な情報を優先的に提示することで小画面デバイス（Pervasive device）ユーザや、拡大ブラウザを用いている弱視ユーザにも読みやすいページに変換する技術が研究されている。IBM WebSphere Transcoding PublisherにもXPathベースのアノテーション記述に基づいてpage clippingを行う機能が実装されている。

【０００７】

このようにウェブ・ページコンテンツの一部分を適切に切り出すことにより、高度に再利用できることが知られている。（１）Webページの部分切り出しの従来技術の方法としては次の（ａ）XPathを用いる方法及び（ｂ）独自タグを用いる方法の二つの方法がある。

【０００８】

（ａ）XPathを用いる方法：

XPathを用いる方法はstaticで変化しないことが保証されている場合には強力な手法である。例えば、非特許文献１では、携帯端末用のページを生成するためにXpath指定によるコンテンツの切り出しが実装されている。しかし、指定の煩雑さ、応用範囲の狭さ等から、実際には携帯端末用の別ページが用意される場合が多く、広まっていない現実がある。また、非特許文献２では、Webページの一部分を選択し、その中で入力部分と出力部分を選択することでウェブ・ページを容易にWEBサービス化可能な枠組みを提案している。この技術はGUI環境で容易に切り出し、サービスの結合を行える点で優れているが、切り出しに関してはXPathに依存しているという問題がある。さらに、非特許文献３では、IBMのトップ・ページ等からXPathによって画像や記事のリストを切り出し、「パーソナルニュースペーパー」の一部に組み込んでいる。レイアウト変更によって切り出し部分がずれてしまうため、XPathの定義ファイルを人手で修正した上で自動配信することで対処している。

【０００９】

（ｂ） 独自タグを用いる方法：

該方法では、HTMLタグの中に独自のタグを混ぜる。HTMLコメントに特別な文字列を指定することもある。LYCOS、YAHOO等のポータルサービスで広く用いられている。例えば、ショッピング・ページのお勧め商品に関する説明をトップ・ページにも表示するといった用途でこの手法が用いられている。この手法は簡易HTMLパーザ（parser）等によって処理できるため、HTMLパーザを用いる場合に用いられることが多い。該方法では、元のコンテンツを変更しなければならないという問題がある。

【００１０】

ウェブ・ページコンテンツの一部分切り出し技術ではないが、本発明に類似する従来技術を列挙する。

（２）XPathセットを手がかりとしたダイナミック・アノテーション・マッチング方法(特願２００１－３３３２６０。ただし、本願明細書作成時ではまだ出願公開されていない。）

該方法では、アノテーション内に含まれるXPathを手がかりとして、複数のアノテーションの候補から適切なものを選択する。該方法により、すべてのレイアウトをカバーするだけのアノテーションを用意することで多くの場合正しいアノテーション・マッチングを行うことができるようになった。しかし、オーサリングの段階でXpathが誤ったノードを指すことも多く、これを修正するための機能としてサイト・パターン・アナライザが持っている、空コンテンツ・アラート、漏れテキスト・アラート、XPathの半自動修正等の機能が開発されたが、調整作業には手間がかかるのが現状である。

【００１１】

（３）その他のアノテーション・マッチング方法：

RDF等多くの場合、アノテーションは対照表かURLの正規表現を利用してアノテーションとページのマッチングを行っている。これらの手法とはダイナミックなコンテンツでのマッチングを行っている点で大きく異なっている。

【００１２】

（４）差分演算とその利用：

差分演算を用いて、アップデートされた情報のみを提示及び再利用したり、notificationメールを送信等したりするサービス・技術としては、DiffWeb （例：非特許文献４）、 HTML Diff（例：非特許文献５）及びMindIt（例：非特許文献６）等が知られている。これらの技術は「一つ前の過去のページ」との差分演算をおこない、そこから取り出したコンテンツを利用している。これに対し、本発明では目的が「マッチング・パターンの生成」である点で大きく異なっている。また要素技術にしても複数バージョンの過去ページとの差分演算および統計処理、隣接ページの概念とその差分演算等大きく異なっている。

【００１３】

（５）差分演算によるシンプリフィケーション技術(特許文献１)：

該技術では、差分演算によって一つのページからページ独自の情報を取り出して単純化する。該技術は、隣接ページのリストアップとその差分演算を行う点で共通性があるものの、ウェブ・ページコンテンツの一部切り出しについての具体的な方法を示唆しない。

【００１４】

（６）ツリー構造のマッチング技術：

ツリー構造を対象に、その構造によるマッチング技術としては、正規表現マッチング技術(TRex)，生垣オートマトンに基づくツリー構造のマッチングとスキーマ言語への応用(relax，relaxNG)等が研究されている。これらの技術はマッチング・パターンの存在を前提としてマッチするサブツリー(ノード)の探索を行う技術であり、マッチング・パターンの自動生成への関連を示唆しない。

【００１５】

（７）マッチング・パターンの自動生成についての関連した技術：

XMLのサンプル群からそれらにマッチするスキーマ記述を自動生成する「Examplotron」がある。この技術はXMLファイル群からある種のマッチング・パターンを自動生成する点で類似しているが、対象が「整形式（well-formatted）」な「ある暗黙のスキーマにのっとった」XMLファイル群である点、さらにタグの「入れ子構造」を手がかりに厳密なマッチング・パターンを生成する点で、後述の本

発明の構成とは異なっている。

【００１６】

（８）アノテーション付与の作業効率化（特許文献２）：

レイアウト構造が近いページ・ファイルに対して共通のアノテーションを付与することにより、アノテーション付与の作業効率化を図っている。レイアウト構造が近いか否かの判定は、構造記述式の対比に基づいて行われ、ノードの出現態様や出現頻度に係る統計情報に基づくマッチング・パターンは利用しない。

【００１７】

【特許文献１】

特開２００２－５５８７２

【特許文献２】

特開２００２－２４５０６８

【非特許文献１】

WTP (WebSphere Transcoding Publisher, http://www-6.ibm.com/jp/software/network/transcoding/)

【非特許文献２】

CHIP[1]　伊藤 ”GUI部品とWEBサービスの統合による分散アプリケーションの構築手法”，ソフトウェア科学会WISS 2001プロシーディングス（http://ca.meme.hokudai.ac.jp/people/itok/CHIP/indexJ.html)

【非特許文献３】

IBM mySiteOutliner(http://www-6.ibm.com/jp/pc/clubibm/msol/index.shtml)

【非特許文献４】

DiffWeb (http://www.diffweb.com/)

【非特許文献５】

HTML Diff (http://www-db.stanford.edu/c3/c3.html)

【非特許文献６】

MindIt (http://mindit.netmind.com/mindit.shtml)

【００１８】

【発明が解決しようとする課題】

本発明の目的は、ネットワークを介して配信される構造化・階層化コンテンツについて、例えばその一部切り出し及び共通のアノテーションの使い回し等の処理を行う際に、大きな威力を発揮する装置、方法及びプログラムを提供することである。

本発明の他の目的は、XPathを用いたりタグを付加したりすることなく、例えば構造化・階層化コンテンツの一部切り出し及び共通のアノテーションの使い回し等を達成できる構造化・階層化コンテンツ用処理装置、構造化・階層化コンテンツ用処理方法及び構造化・階層化コンテンツ用処理プログラムを提供することである。

【００１９】

【課題を解決するための手段】

本発明では、コンテンツの一部切り出し及び複数のコンテンツに対する共通アノテーションの使い回し等の処理対象としての構造化・階層化コンテンツであるか否かを同定する（identify）ために、XPathではなく、マッチング・パターンを使用する。

【００２０】

本発明では、ターゲット・コンテンツに対する過去及び／又は隣接の構造化・階層化コンテンツを調べ、ターゲット・サブツリーにおける各ノードについての出現態様及び該出現態様の出現頻度に係る統計情報に基づいて各ノードを分類して、マッチング・パターンを生成する。

【００２１】

本発明の構造化・階層化コンテンツ用処理装置では、ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う。さらに、構造化・階層化コンテンツ用処理装置は、マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定

手段、前記ターゲット・コンテンツに対する過去の複数個の構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出手段、過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成手段、前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類手段、及び該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成手段、を有している。

【００２２】

本発明の構造化・階層化コンテンツ用処理方法では、ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う。さらに、本発明の構造化・階層化コンテンツ用処理方法は、マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定ステップ、前記ターゲット・コンテンツに対する過去の複数個の構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出ステップ、過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成ステップ、前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類ステップ、及び該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成ステップ、を有している。

【００２３】

過去の各構造化・階層化コンテンツの代わりにターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを利用することもできる。ネットワークには、インターネットの外、イントラネット、エクストラネット等が含まれる。構造化・階層化コンテンツとは、コンテンツ本体の他に、構造情報及び階層情報を含むコンテンツと定義する。構造化・階層化コンテンツとして、例えばXML文書及びウェブ・ページ（HTMLファイル）がある。

【００２４】

本発明の構造化・階層化コンテンツ用処理プログラムは、前記構造化・階層化コンテンツ用処理方法の各ステップをコンピュータに実行させる。

【００２５】

判定対象の構造化・階層化コンテンツがターゲット・コンテンツに対して隣接構造化・階層化コンテンツであるか否かは、ＵＲＬ及び／又はレイアウトの近似性が判定要素とされる。デフォルト状態では、システムが、両者の近似性を加味して、総合的な近似性、すなわち判定対象の構造化・階層化コンテンツがターゲット・コンテンツに対して隣接構造化・階層化コンテンツであるか否かを判定する。このようなデフォルトに対して、オーサは、具体的近似性、すなわち判定対象の構造化・階層化コンテンツのＵＲＬ及び／又はレイアウトが具体的にどのようになっていれば、判定対象の構造化・階層化コンテンツがターゲット・コンテンツに対して隣接構造化・階層化コンテンツであると判定するかの具体的条件を、各ターゲット・コンテンツの具体的内容に基づいて定め、該具体的条件をデフォルトに代えて、コンピュータに指示することも可能である。構造化・階層化コンテンツがターゲット・コンテンツに対して隣接構造化・階層化コンテンツであるか否かの判定を行う各手段（例：出現態様検出手段及び統計情報生成手段）及び各ステップ（例：出現態様検出ステップ及び統計情報生成ステップ）は、該具体的条件に基づいて判定を実施する。

【００２６】

「隣接構造化・階層化コンテンツ」とは、そのＵＲＬがターゲット・コンテンツのＵＲＬとは相違するものの、（ａ）そのＵＲＬがターゲット・コンテンツの

ＵＲＬとの同一部分を所定割合以上で有している構造化・階層化コンテンツ、及び／又は（ｂ）そのレイアウトの少なくとも主要部がターゲット・コンテンツのレイアウトと同一である構造化・階層化コンテンツであると、定義できる。（ｂ）で定義される隣接構造化・階層化コンテンツには、そのレイアウトがターゲット・コンテンツのレイアウトとの同一の領域を所定割合以上、有している構造化・階層化コンテンツを含むものとする。

【００２７】

隣接構造化・階層化コンテンツは少なくとも次の（ａ）及び（ｂ）のものを含む。

（ａ）属しているディレクトリがターゲット・コンテンツと共通である構造化・階層化コンテンツ。構造化・階層化コンテンツがウェブ・コンテンツである場合の具体例（asahi.com）は次の通りである。

ターゲット・コンテンツとしてウェブ・コンテンツのＵＲＬ：

http://www.asahi.com/0606/news/national06015.html

に対する隣接構造化・階層化コンテンツとして例えば次のＵＲＬ。

http://www.asahi.com/0606/news/national06012.html

http://www.asahi.com/0606/news/national06013.html

http://www.asahi.com/0606/news/national06014.html

（ｂ）所定の階層数（例えば２階層）上のディレクトリがターゲット・コンテンツのものと共通である構造化・階層化コンテンツ。構造化・階層化コンテンツがウェブ・コンテンツである場合の具体例（cnn.com）は次の通りである。

ターゲット・コンテンツとしてウェブ・コンテンツのＵＲＬ：

http://www.cnn.com/2000/US/06/05/sea.based.defense/index.html

に対する隣接構造化・階層化コンテンツとして例えば次のＵＲＬ。

http://www.cnn.com/2000/US/06/05/dday.remembrance/index.html

http://www.cnn.com/2000/US/06/05/helicopter.escape.03/index.html

http://www.cnn.com/2000/US/06/05/curbing.terrorism.02/index.html

【００２８】

【発明の実施の形態】

構造化・階層化コンテンツ用処理装置は、ウェブ・ページの一部分を切り出す方法として、切り出したい領域を指定するだけで自動的にマッチング・パターンを高い精度で自動的に生成し、適切なコンテンツのロバスト（robust）な切り出しを実現する。マッチング・パターンの生成は「複数ページ（以降、ウェブ・コンテンツを適宜、「ページ」と呼ぶことにする。）との差分統計量」を基にする。指定された領域(DOMツリー上のあるノード）をあらかじめ保存しておいた過去のページ群と比較（差分演算）し、統計量を算出し、定常なノード、更新され必ず存在するノード、追加・消滅するノードに分類する。このようなノードの分類を施した上で繰り返しパターンの検出等の処理を行ったサブツリーがアノテーションのマッチング・パターンになる。過去のページが存在しない場合、隣接ページと同様の処理を行うことで同様にマッチング・パターンを得る。このようなマッチング・パターンは従来のXPathや埋め込みタグに基づく手法とは異なり、オリジナルのコンテンツを変更する必要がなく、マッチング・パターンを外部アノテーションとして適用するだけで正確な切り出しが可能になる。さらに、上位ノードの変更があってもまったく影響しない点で格段にロバストである。

【００２９】

「アノテーション」とは、所定の構造化・階層化コンテンツＡから別の構造化・階層化コンテンツＢを作成するときに、Ｂに付加された所定の情報のことを言うものとする。この付加的な所定情報には、（ａ）コンテンツＡの一部分を指定する情報、（ｂ）コンテンツＡにおいて指定された部分に関する情報、及び／又は（ｃ）上記（ａ）及び（ｂ）を適宜組み合わせた情報を含む。Ｂの具体例を挙げると、画面表示態様のＢでは、画面表示態様のＡの下側にＡの主要項目をまとめたリストや、フォント・サイズ変更等の各種指示リストが付加される。このように付加されたものがアノテーションであり、ユーザは、該付加部の主要項目リストの項目をクリックすれば、Ｂ内のＡ部分の対応個所へジャンプできるようになっていたり、各種指示リストの項目をクリックすれば、Ａ部分を含むＢの字が大きく表示される等の対応の処理が行われたりする。なお、マッチング・パターンは、それをコンテンツＡの一部分を指定する情報として利用し、付加情報（そのコンテンツ部分の役割、重要度などの情報）と組み合わせることによりアノテ

ーションとして機能させることができる。

【００３０】

図１はウェブ・コンテンツ処理装置１４を装備する構造化・階層化コンテンツ用処理システム１０の構成図ある。本発明が適用されるネットワークは、インターネット１２に限定されず、イントラネット及びエクストラネット等であってもよい。ウェブ・コンテンツ処理装置１４、ウェブ・クライアント１５及びウェブ・サーバ１６は、インターネット１２へ接続され、インターネット１２を介して相互にデータを送受可能になっている。１個のウェブ・コンテンツ処理装置１４は、構造化・階層化コンテンツ用処理装置として振る舞い、複数個のウェブ・クライアント１５からの要求に応じて複数個のウェブ・サーバ１６の中から対応する１個又は複数個のウェブ・サーバ１６よりウェブ・コンテンツをHTTP(HyperText Transfer Protocol)により取り寄せ、該ウェブ・コンテンツに所定の処理、例えばアノテーション付与及び／又はコンテンツ切り出し等の処理を行って、ウェブ・クライアント１５へ処理済みのウェブ・コンテンツを送信する。なお、ユーザが実際に操作するウェブ・クライアント１５としてのパーソナル・コンピュータは、インターネット１２へ直接、接続されていなくてよい。該パーソナル・コンピュータは、直接的には社内のＬＡＮへ接続され、該ＬＡＮのプロキシ・サーバやルータを介してインターネット１２へ接続されていてもよい。

【００３１】

図２は構造化・階層化コンテンツ用処理装置１８のブロック図である。構造化・階層化コンテンツ用処理装置１８は、それが処理対象とする構造化・階層化コンテンツがウェブ・コンテンツである場合には、図１のウェブ・コンテンツ処理装置１４となる。構造化・階層化コンテンツ用処理装置１８のオーサ（Author）は、複数の構造化・階層化コンテンツ（例えばウェブ・コンテンツ）に共通に使用できるアノテーションを作成したり、１個又は複数個の構造化・階層化コンテンツから所定のコンテンツ部分を切り出して（ここで言う「切り出し」とは切り出し元の構造化・階層化コンテンツから切り出しコンテンツ部分が削除されることを意味せず、該切り出しコンテンツ部分は切り出し元の構造化・階層化コンテンツに残る。∴ここで言う「切り出し」とは厳密に言うと「コピー」である。）

、切り出した１個又は複数個のコンテンツ部分を貼り付けて新規な構造化・階層化コンテンツを作成したり等の、構造化・階層化コンテンツ編集作業を行う。オーサは、マッチング・パターンを抽出しようとする構造化・階層化コンテンツとしてのターゲット・コンテンツ２０をネットワークを介して所定の構造化・階層化コンテンツ・サーバから読み込む。オーサは、次に、ターゲット・コンテンツ２０から所定のコンテンツ部分を指定する。該指定されたコンテンツ部分を「ターゲット・コンテンツ部分２１」と呼ぶことにする。構造化・階層化コンテンツ用処理装置１８は、ターゲット・コンテンツ部分２１に対して、ターゲット・コンテンツ２０のDOMツリー上で、ターゲット・コンテンツ部分２１を含む範囲に係るサブツリーをターゲット・サブツリーとして自動的に設定する。ターゲット・サブツリーは、ターゲット・コンテンツ部分２１を含む範囲に係ることが要件であり、該範囲は、必要最小限の範囲にすることが好ましく、ターゲット・コンテンツ部分２１より適当に大きいコンテンツ部分の範囲に設定されてもよい。オーサは、今回の編集作業に先立ち、構造化・階層化コンテンツ・データベース２６に対して、ターゲット・コンテンツ２０のXPathを予め（例えば、今回の編集作業の１週間前、１０日前、１月前等）通知しておく。構造化・階層化コンテンツ・データベース２６は、通知後、自動的かつ定期的にターゲット・コンテンツ２０に係るコンテンツにアクセスして、該コンテンツを蓄積する。したがって、該ターゲット・コンテンツ２０の今回のユーザ作業では、ターゲット・コンテンツ２０に対する過去の構造化・階層化コンテンツが十分な数だけ構造化・階層化コンテンツ・データベース２６に蓄積されている。出現態様検出手段２７は、ターゲット・コンテンツ２０に対する過去の構造化・階層化コンテンツを構造化・階層化コンテンツ・データベース２６より１個ずつ又はまとめて読み出し、ターゲット・コンテンツ部分２１に係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照して、ターゲット・サブツリーの各ノードの出現態様を検出する。ターゲット・コンテンツ２０に対する過去の複数個の構造化・階層化コンテンツは、好ましくは、現在の時点、すなわちマッチング・パターン生成処理時点に対して過去所定期間内の構造化・階層化コンテンツである。なお、ターゲット・コンテンツ２０とターゲット・コンテンツ２０に対して

過去の構造化・階層化コンテンツとは、URI(Uniform Resource Locator)が同一となっている。統計情報生成手段２８は、過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する。分類手段２９は、出現態様検出手段２７における出現態様検出結果及び統計情報生成手段２８が生成した統計情報に基づいてターゲット・サブツリーの各ノードを分類する。

【００３２】

出現態様検出手段２７、統計情報生成手段２８及び分類手段２９における処理をより具体的に説明する。出現態様検出手段２７では、ターゲット・コンテンツ２０に係るターゲット・サブツリーと過去の１個の構造化・階層化コンテンツのツリーとを対照することにより、ターゲット・サブツリーの各ノードについて（Ｎ１）構造化・階層化コンテンツにも出現しかつ内容が同一であるノード、（Ｎ２）構造化・階層化コンテンツにも出現するが内容が異なるノード、（Ｎ３）構造化・階層化コンテンツには出現しないノードのいずれであるかを区分けできる。なお、ノードの内容とは、構造化・階層化コンテンツとしてのXMLでは、開始タグと終了タグとの間の記述内容のことである。出現態様検出手段２７が過去の複数の所定個数の構造化・階層化コンテンツの各々についてそのツリーをターゲット・サブツリーと対照することにより、ターゲット・サブツリーの各ノードについて、（Ｎ１）及び（Ｎ２）の出現頻度の統計情報を検出できる。統計情報生成手段２８はこの統計情報を生成する。分類手段２９は、（Ｎ１）及び（Ｎ２）の態様で出現する頻度について予め設定した閾値Ｖ１，Ｖ２をもつ。典型的にはＶ１＝Ｖ２であるが、Ｖ１及びＶ２は相互に異なった値であってもよい。典型的にはＶ１＝Ｖ２＝７０％とする。分類手段２９におけるノード分類の具体例は次の通りである。（Ｎ１）の態様による出現頻度≧Ｖ１であるノードは定常ノードに分類される。（Ｎ２）の態様による出現頻度≧Ｖ２であるノードは更新ノードに分類される。定常ノード及び更新ノードのいずれにも分類されなかったノードは付加ノードに分類される。

【００３３】

マッチング・パターン生成手段３０は分類手段２９における分類結果に基づい

てマッチング・パターンを生成する。マッチング・パターン生成手段３０において生成されたマッチング・パターンとコンテンツ部分とのマッチング処理の詳細は後述の図６において説明する。

【００３４】

図３はマッチング・パターン生成手段３０のより具体的なブロック図である。繰り返し部分検出手段３４は、定常ノード、更新ノード及び付加ノードの分類に基づいてターゲット・サブツリーにおける繰り返し部分を検出する。繰り返し情報付きマッチング・パターン生成手段３５は、該繰り返し部分の存在情報を含むマッチング・パターンを生成する。こうして、生成されたマッチング・パターンは、マッチするか否かを判定される構造化・階層化コンテンツが、繰り返し部分を任意の回数、繰り返すものであっても、該マッチング・パターンにマッチするものとして使用可能となる。

【００３５】

図４は分類手段２９のより具体的なブロック図である。構造化・階層化コンテンツは表示時のレイアウトを良好にするために、スペーサ用イメージ及びビュレット・イメージを含むことがある。スペーサ用イメージとは、HTMLファイルの「spacer GIF」に対応し、空白領域を確保するために、１個の構造化・階層化コンテンツに複数個、使用され、それぞれ指定サイズの異なるイメージである。これに対し、ビュレット・イメージ（コンテンツにおいて列記された各項目の先頭に置くマーク）とは、HTMLファイルの「bullet イメージ」に対応し、１個の構造化・階層化コンテンツに複数個、使用され、サイズは、同一を指定されているか、又は指定無しとされている。スペーサ用イメージ検出手段３８は、ターゲット・サブツリーのノードについてそれがスペーサ用イメージに係るノードであるか否かを検出する。ビュレット・イメージ検出手段３９は、ターゲット・サブツリーのノードについてそれがビュレット・イメージに係るノードであるか否かを検出する。第１の分類付け手段４０は、スペーサ用イメージに係るノードは付加ノードと分類する。第２の分類付け手段４１は、ビュレット・イメージに係るノード同士へは、その表示内容が異なっていても定常ノード、更新ノード又は付加ノードの同一分類に割り当てる。分類出力手段４２は、第１の分類付け手段４０及

び第２の分類付け手段４１によるノードの分類付けをまとめる機能を備え、分類手段２９の出力を生成する。

【００３６】

図２の構造化・階層化コンテンツ用処理装置１８はターゲット・コンテンツに対する過去の構造化・階層化コンテンツ（ターゲット・コンテンツに対してURIが同一となっている過去の構造化・階層化コンテンツ）に基づいてマッチング・パターンを生成するが、ターゲット・コンテンツに対する隣接の構造化・階層化コンテンツに基づいてマッチング・パターンを生成することもできる。隣接の構造化・階層化コンテンツに基づくマッチング・パターンの生成は、（ａ）ターゲット・コンテンツに対する過去のコンテンツ部分がないときのみ実施されてもよいし、（ｂ）ターゲット・コンテンツに対する過去のコンテンツ部分の有無に関係なく実施されてもよい。例えば朝日新聞(www.asahi.com)のビジネス記事ページは次のようにＵＲＬの中に日付が含まれいて、現在の現在を含む所定期間、最新のビジネス記事と共に閲覧可能になっている。なお、下記の例では、該ビジネス記事は１０月１９日のものである。

「http://www.asahi.com/business/update/1019/002.html」

このようなケースに対しても、適切なマッチング・パターンを生成するため、本発明では「ターゲット・コンテンツに対する隣接構造化・階層化コンテンツ」なる概念を導入する。隣接構造化・階層化コンテンツとは、ターゲット・コンテンツに対して近似したURIを有し、マッチング・パターンによるマッチング判定のときに構造化・階層化コンテンツと同一グループに属させる構造化・階層化コンテンツである。URIの近似範囲は、オーサがどの程度の相違以下を同一グループに属すると判断するかにより変動する。URIには、各階層のディレクトリ（朝日新聞ビジネス記事の例では、/で区切られている部分）が含まれるが、隣接構造化・階層化コンテンツか否かの判定対象となっているコンテンツのURIが、ターゲット・コンテンツのURIに対して最高位の階層から所定数（１以上の数）の階層までのディレクトリは同一で、該同一ディレクトリの階層より下位の階層のディレクトリのみが相違しているときは、該判定対象のコンテンツ部分は隣接コンテンツ部分と判定してもよい。隣接コンテンツ部分の具体例を列挙すると、次の

通りである。次の場合には、判定対象の構造化・階層化コンテンツは隣接構造化・階層化コンテンツと判定される。

（ａ）URIにおいて日付と認められる部分のみがターゲット・コンテンツに対して相違している。前述の朝日新聞ビジネス記事の例では、”1019”である。

（ｂ）URIにおいて番号付けとして使用されている部分のみがターゲット・コンテンツに対して相違している。前述の朝日新聞ビジネス記事の例では、”002.html”である。

（ｃ）前述の（ａ）及び（ｂ）のみがターゲット・コンテンツに対して相違している。

【００３７】

図２の構造化・階層化コンテンツ用処理装置１８が過去の構造化・階層化コンテンツに代えて隣接構造化・階層化コンテンツに基づいてマッチング・パターンを生成する場合について、過去の構造化・階層化コンテンツに基づいてマッチング・パターンを生成する場合との相違点のみを説明する。構造化・階層化コンテンツ・データベース２６は、任意の構造化・階層化コンテンツがオーサにより今回のターゲット・コンテンツ２０として選択されるのに対処して、所定の構造化・階層化コンテンツに対する複数個の隣接構造化・階層化コンテンツを予め蓄積する。出現態様検出手段２７は、ターゲット・コンテンツ２０に対する隣接構造化・階層化コンテンツを構造化・階層化コンテンツ・データベース２６より１個ずつ又はまとめて読み出し、ターゲット・コンテンツ２０に係るターゲット・サブツリーとその隣接の各構造化・階層化コンテンツに係るツリーとを対照して、ターゲット・サブツリーの各ノードの出現態様を検出する。統計情報生成手段２８は、複数個の隣接構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する。分類手段２９は、出現態様検出手段２７における出現態様検出結果及び統計情報生成手段２８が生成した統計情報に基づいてターゲット・サブツリーの各ノードを分類する。過去の構造化・階層化コンテンツに代えて隣接構造化・階層化コンテンツを使用する場合の出現態様検出手段２７、統計情報生成手段２８及び分類手段２９における処理をより具体的に説明すると、次の通りである。出現態様

検出手段２７では、ターゲット・コンテンツ２０に係るターゲット・サブツリーと１個の隣接構造化・階層化コンテンツのツリーとを対照することにより、ターゲット・サブツリーの各ノードについて（Ｎ１）構造化・階層化コンテンツにも出現しかつ内容が同一であるノード、（Ｎ２）構造化・階層化コンテンツにも出現するが内容が異なるノード、（Ｎ３）構造化・階層化コンテンツには出現しないノードのいずれであるかを区分けできる。出現態様検出手段２７が複数の所定個数の隣接構造化・階層化コンテンツの各々についてそのツリーをターゲット・サブツリーと対照することにより、ターゲット・サブツリーの各ノードについて、（Ｎ１）及び（Ｎ２）の出現頻度の統計情報を検出できる。統計情報生成手段２８はこの統計情報を生成する。分類手段２９は、（Ｎ１）及び（Ｎ２）の態様で出現する頻度について予め設定した閾値Ｖ１，Ｖ２をもつ。典型的にはＶ１＝Ｖ２であるが、Ｖ１及びＶ２は相互に異なった値であってもよい。典型的にはＶ１＝Ｖ２＝７０％とする。分類手段２９におけるノード分類の具体例は次の通りである。（Ｎ１）の態様による出現頻度≧Ｖ１であるノードは定常ノードに分類される。（Ｎ２）の態様による出現頻度≧Ｖ２であるノードは更新ノードに分類される。定常ノード及び更新ノードのいずれにも分類されなかったノードは付加ノードに分類される。

【００３８】

なお、図３のマッチング・パターン生成手段３０及び図４の分類手段２９は、過去の構造化・階層化コンテンツに代えて隣接構造化・階層化コンテンツに基づいてマッチング・パターンを生成する場合にも適用される。

【００３９】

図５は過去の構造化・階層化コンテンツに基づいてマッチング・パターンを生成する方法のフローチャートである。該マッチング・パターン生成方法の各ステップの動作主体は、該マッチング・パターン生成方法の各ステップを実行するプログラムをインストールされるコンピュータ（該コンピュータは図１の例ではウェブ・コンテンツ処理装置１４に相当する。）である。Ｓ４６では、ターゲット・サブツリーを設定する。オーサは、マッチング・パターンを抽出しようとする構造化・階層化コンテンツとしてのターゲット・コンテンツ２０をネットワーク

を介して所定の構造化・階層化コンテンツ・サーバから読み込み、次に、ターゲット・コンテンツ２０から所定のコンテンツ部分を指定する。Ｓ４６では、ターゲット・コンテンツ部分２１に対して、ターゲット・コンテンツ２０のDOMツリー上で、ターゲット・コンテンツ部分２１の範囲を含むサブツリーをターゲット・サブツリーとして自動的に設定する。ターゲット・サブツリーは、ターゲット・コンテンツ部分２１を含む範囲に係ることが要件であり、該範囲は、必要最小限の範囲にすることが好ましく、ターゲット・コンテンツ部分２１より適当に大きいコンテンツ部分の範囲に設定されてもよい。Ｓ４７では、ターゲット・コンテンツ２０に対する過去の構造化・階層化コンテンツを構造化・階層化コンテンツ・データベース２６より１個ずつ又はまとめて読み出す。Ｓ４８では、ターゲット・コンテンツ２０に係るターゲット・サブツリーとその過去の各構造化・階層化コンテンツに係るツリーとを対照して、ターゲット・サブツリーの各ノードの出現態様を検出する。ターゲット・コンテンツ２０に対する過去の複数個の構造化・階層化コンテンツは、好ましくは、現在の時点、すなわちマッチング・パターン生成処理時点に対して過去所定期間内の構造化・階層化コンテンツである。なお、ターゲット・コンテンツ２０とターゲット・コンテンツ２０に対して過去の構造化・階層化コンテンツとは、URI(Uniform Resource Locator)が同一となっている。Ｓ４９では、過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する。Ｓ５０では、出現態様検出手段２７における出現態様検出結果及び統計情報生成手段２８が生成した統計情報に基づいてターゲット・サブツリーの各ノードを分類する。

【００４０】

Ｓ４８、Ｓ４９及びＳ５０における処理をより具体的に説明する。Ｓ４８では、ターゲット・コンテンツ２０に係るターゲット・サブツリーと過去の１個の構造化・階層化コンテンツのツリーとを対照することにより、ターゲット・サブツリーの各ノードについて（Ｎ１）構造化・階層化コンテンツにも出現しかつ内容が同一であるノード、（Ｎ２）構造化・階層化コンテンツにも出現するが内容が異なるノード、（Ｎ３）構造化・階層化コンテンツには出現しないノードのいず

れであるかを区分けできる。Ｓ４８が過去の複数の所定個数の構造化・階層化コンテンツの各々についてそのツリーをターゲット・サブツリーと対照することにより、ターゲット・サブツリーの各ノードについて、（Ｎ１）及び（Ｎ２）の出現頻度の統計情報を検出できる。Ｓ４９はこの統計情報を生成する。Ｓ５０は、（Ｎ１）及び（Ｎ２）の態様で出現する頻度について予め設定した閾値Ｖ１，Ｖ２をもつ。典型的にはＶ１＝Ｖ２であるが、Ｖ１及びＶ２は相互に異なった値であってもよい。典型的にはＶ１＝Ｖ２＝７０％とする。Ｓ５０におけるノード分類の具体例は次の通りである。（Ｎ１）の態様による出現頻度≧Ｖ１であるノードは定常ノードに分類される。（Ｎ２）の態様による出現頻度≧Ｖ２であるノードは更新ノードに分類される。定常ノード及び更新ノードのいずれにも分類されなかったノードは付加ノードに分類される。

【００４１】

Ｓ５１では、Ｓ５０における分類結果に基づいてマッチング・パターンを生成する。図６は図５のマッチング・パターン生成方法において生成されたマッチング・パターンを使用するマッチング判定方法のフローチャートである。Ｓ５５では、これからマッチング・パターンとのマッチングを判定しようとするコンテンツ部分（以下、「被判定コンテンツ部分」と言う。）を読み出す。Ｓ５６では、被判定コンテンツ部分がマッチング・パターンとマッチする部分をもつか否かを判定する。マッチング・パターンとマッチすると判定されるときの被判定コンテンツ部分は、該被判定コンテンツ部分を含む構造化・階層化コンテンツ（以下、「被判定コンテンツ」と言う。）において任意の位置にあってよい。すなわち、マッチング・パターンとマッチする被判定コンテンツ部分は、被判定コンテンツの任意の位置にあっても、マッチング・パターンとマッチすると、正しく判定される。Ｓ５６の判定が正であれば、Ｓ５７へ進み、否であれば、該方法を終了する。Ｓ５７では、被判定コンテンツ部分に対して所定の処理を実施する。該所定の処理とは、例えば、（ａ）被判定コンテンツのコンテンツ部分への関連情報の関連付け、（ｂ）被判定コンテンツのコンテンツ部分を他の構造化・階層化コンテンツに利用するために該被判定コンテンツの該被判定コンテンツ部分をコピーする処理（当業者は、該処理を「切り出し」と呼んでいる。）である。（ａ）の

関連情報とは例えばアノテーションである。

【００４２】

図７は図５のマッチング・パターン生成ステップ（Ｓ５１）をより具体的に示すフローチャート部分である。Ｓ６０では、定常ノード、更新ノード及び付加ノードの分類に基づいてターゲット・サブツリーにおける繰り返し部分を検出する。Ｓ６１では、Ｓ６０において検出した繰り返し部分の存在情報を含むマッチング・パターンを生成する。こうして、生成されたマッチング・パターンは、マッチするか否かを判定される構造化・階層化コンテンツが、任意の回数の繰り返し部分をもっていても、該マッチング・パターンにマッチするものとして使用可能となる。

【００４３】

図８は分類手段２９のより具体的なブロック図である。図８では、Ｓ６４及びＳ６５の系列と、Ｓ６６及びＳ６７の系列とは並列処理されるように記載されているが、一方及び他方の系列をそれぞれ先行及び後続させる直列処理にしてもよい。Ｓ６４では、ターゲット・サブツリーのノードについてそれがスペーサ用イメージに係るノードであるか否かを検出する。Ｓ６５では、スペーサ用イメージに係るノードは付加ノードと分類する。Ｓ６６では、ターゲット・サブツリーのノードについてそれがビュレット・イメージに係るノードであるか否かを検出する。Ｓ６７では、ビュレット・イメージに係るノード同士は、その表示内容が異なっても定常ノード、更新ノード又は付加ノードの同一分類に割り当てる。Ｓ６８では、Ｓ６５及びＳ６７の分類結果をまとめ、出力する。

【００４４】

図９はターゲット・コンテンツに対して隣接する複数個の構造化・階層化コンテンツに基づいてマッチング・パターンを生成する方法のフローチャートである。図５において、ターゲット・コンテンツに対する過去の構造化・階層化コンテンツに基づいてマッチング・パターンを生成する方法について説明したが、図９に係る生成方法は、（ａ）ターゲット・コンテンツに対する過去のコンテンツ部分がないときのみ実施されてもよいし、（ｂ）ターゲット・コンテンツに対する過去のコンテンツ部分の有無に関係なく実施されてもよい。図５のフローチャー

トに対する図９のフローチャートの相違点は、図５のＳ４７～Ｓ５０に代えて、Ｓ４７ｂ～Ｓ５０ｂを実施することである。相違点のみ説明する。

【００４５】

Ｓ４７ｂでは、ターゲット・コンテンツ２０に対する隣接構造化・階層化コンテンツを構造化・階層化コンテンツ・データベース２６より１個ずつ又はまとめて読み出す。Ｓ４８ｂでは、ターゲット・コンテンツ２０に係るターゲット・サブツリーとその各隣接構造化・階層化コンテンツに係るツリーとを対照して、ターゲット・サブツリーの各ノードの出現態様を検出する。Ｓ４９ｂでは、複数個の隣接構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する。Ｓ５０ｂでは、出現態様検出手段２７における出現態様検出結果及び統計情報生成手段２８が生成した統計情報に基づいてターゲット・サブツリーの各ノードを分類する。Ｓ４８ｂ、Ｓ４９ｂ及びＳ５０ｂにおける処理をより具体的に説明する。Ｓ４８ｂでは、ターゲット・コンテンツ２０に係るターゲット・サブツリーと隣接の１個の構造化・階層化コンテンツのツリーとを対照することにより、ターゲット・サブツリーの各ノードについて（Ｎ１）構造化・階層化コンテンツにも出現しかつ内容が同一であるノード、（Ｎ２）構造化・階層化コンテンツにも出現するが内容が異なるノード、（Ｎ３）構造化・階層化コンテンツには出現しないノードのいずれであるかを区分けできる。Ｓ４８ｂが隣接の複数の所定個数の構造化・階層化コンテンツの各々についてそのツリーをターゲット・サブツリーと対照することにより、ターゲット・サブツリーの各ノードについて、（Ｎ１）及び（Ｎ２）の出現頻度の統計情報を検出できる。Ｓ４９ｂではこの統計情報を生成する。Ｓ５０ｂでは、（Ｎ１）及び（Ｎ２）の態様で出現する頻度について予め設定した閾値Ｖ１，Ｖ２を取得する。典型的にはＶ１＝Ｖ２であるが、Ｖ１及びＶ２は相互に異なった値であってもよい。典型的にはＶ１＝Ｖ２＝７０％とする。Ｓ５０ｂにおけるノード分類の具体例は次の通りである。（Ｎ１）の態様による出現頻度≧Ｖ１であるノードは定常ノードに分類される。（Ｎ２）の態様による出現頻度≧Ｖ２であるノードは更新ノードに分類される。定常ノード及び更新ノードのいずれにも分類されなかったノードは付加ノードに分類される。Ｓ５１では

、Ｓ５０における分類結果に基づいてマッチング・パターンを生成する。

【００４６】

なお、図７及び図８のフローチャートは、過去の構造化・階層化コンテンツに代えて隣接構造化・階層化コンテンツに基づいてマッチング・パターンを生成する場合にも適用される。

【００４７】

【実施例】

実施例は構造化・階層化コンテンツとしてウェブ・コンテンツを選択したものである。過去ページ及び隣接ページとの差分演算の結果を用いて統計的に算出されるコンテンツのマッチング・パターンを切り出し部分特定に用いる。図１０はウェブ・コンテンツ用処理装置７４の構成図である。ウェブ・クライアント７６、トランスコーディング・モジュール７７及びウェブ・サーバ７８はインターネットへ接続され、相互にデータを送受自在になっている。ユーザ７５は、ウェブ・クライアント７６を操作して、トランスコーディング・モジュール７７へトランスコーディッドHTML８１の送付を要求する。トランスコーディング・モジュール７７は、ウェブ・クライアント７６からの要求を受付けると、対応のウェブ・サーバ７８からターゲットHTML７９を受け取り、ターゲットHTML７９をアノテーション・データベースからのアノテーションに基づいて加工し（ｔｒａｎｓｃｏｄｅ）、トランスコーディッドHTML８１をウェブ・クライアント７６へ送る。なお、アノテーション・データベースは、典型的にはトランスコーディング・モジュール７７の実装されているコンピュータに装備されているが、トランスコーディング・モジュール７７とは別の場所にあって、インターネットを介してトランスコーディング・モジュール７７へ接続されていてもよい。アノテーション・エディタ８５、キャッシュ・データベース８６及びサイト・パターン・アナライザ８８は、アノテーション・データベースを装備するコンピュータに実装又は装備される。キャッシュ・データベース８６には、隣接ページの算出アルゴリズム、複数バージョンの過去ページをキャッシュする仕組み、及び指定されたURLを定期的に巡回して該URLのページを取得する機能を装備する。キャッシュ・データベース８６は、アノテーション・エディタ８５を使用して、各ターゲットHTML７

９についてのアノテーションを作成する。アノテーション・オーサ８４の作業効率を向上するために、同一のアノテーションを複数個のターゲットHTML７９に共通に使用するアノテーションの使い回しが行われる。アノテーションの適切な使い回しを達成するために、類似した複数個のターゲットHTML７９同士が１個のグループにまとめられ、各グループには同一のアノテーション・セットが使用される。なお、アノテーション・セットとは、複数個のアノテーションをまとめたものである。ターゲットHTML７９が所定のグループに属するか否かは、ターゲットHTML７９と所定のマッチング・パターンとを対照することにより判定される。

【００４８】

マッチング・パターンは、「ページ内のどの部分に出現してもマッチするアノテーション」を実現するために、利用することが可能である。これによりレイアウトの変更に対してロバストな切り出しを実現できる。以下ではまず基本的な手法である隣接ページと過去ページとの差分によりマッチング・パターンを自動生成する手法を述べてから、実際のユーザインタフェース上でオペレーション例を述べる。

【００４９】

[差分演算に基づく過去ページにおける出現頻度演算]

差分演算は、差分演算によるシンプリフィケーションで用いられた方法と同等のものを前提とする。XMLDiff等、厳密なXMLの差分演算を行うアルゴリズムを用いても本手法は実行可能である。ここでは図１１のようにDOMツリーを一旦シリアライズしてからDPマッチングを用いてLongest common node string(LCNS)を算出する手法を用いるものとする。この手法は正確なツリーの差分演算を行えない代わりに、実用上問題がないことがすでに確認済みであること、高速であること、演算対象のエレメントをコントロールし易いこと等から本手法にも適している。以下の記述では差分演算にこの手法を用いるものとして記述する。また、下記の多くの処理ステップにおいて、差分演算の結果として「共通ノード」を用いている。「共通ノード」とは、２つのDOMツリーに共通するノード群であり、差分演算結果から差分以外の部分を選択することで得ることができる。今回用いるDPマッチングによる差分演算手法では、演算途中で共通部分をLCNSとして得ること

ができるため、実際の差分部分を算出することなく、共通ノードを得ることができる。そのため、演算途中に実際に差分算出は生じないが、一般的には差分演算の変種として捕らえることができるため、以下の記述では「差分演算」という記述を用いる。厳密には「差分演算の途中結果としての共通ノード群（LCNS）」を用いている。

【００５０】

図１１はDPマッチングの概略説明図である。第１及び第２の入力を例えばそれぞれ”KWPSIKAWNA”及び”ABPSAWNDS”とする。DPマッチングにより、それら入力のLongest common node string(LCNS)としての”PSAWN”が出力される。DPマッチングでは、余分な要素（例では第１の入力の”IK”）が割り込んでいても、要素同士の相対順番が同一であれは、それら要素からなるストリングをLCNSとして抽出できる。

【００５１】

図１２は差分演算にDPマッチングを適用した概略説明図である。ターゲット・ページと比較ページ（比較ページは過去ページ又は隣接ページである。）とのDOMツリーのターゲットの部分がそれぞれ直列化（serialise）手段９１，９２へ入力され、ツリー配置から直列配置へ変換される。DPマッチング手段９３は、直列化手段９１，９２からの入力からに基づいてのLongest common node string(LCNS)を算出する。差分手段としてのLCNS除去手段９４は、ターゲット・ページのDOMツリーからLCNSを差し引いた値としての差分DOMツリーを出力する。

【００５２】

ロタイプＡ：過去のページが存在する場合のマッチング・パターンの算出

アノテーション・エディタを用いてアノテーション・オーサがすでにDOMツリー上のあるノード群をすでに指定した状態を考える。

ステップ１： ターゲット・サブツリーを決定する。対象ノード群が共通に持つ祖先ノードを一つ探索する。どんな場合でも<body>ノードは共通に持っているため、このようなノードが必ず存在することは明白である。

ステップ２： キャッシュから過去のページ・リスト取得する。アノテーション・オーサはあらかじめ数日から数週間分の過去のページを保存しておくことが望

ましい。過去のページが多いほどロバストなパターンを生成可能である。

ステップ３： 各過去ページと現在ターゲットになっているページの差分演算を行う(１回目の差分演算)。差分演算のためのシリアライズを行う際には、指定されたグループ内の全エレメントをシリアライズ対象に追加する。DPマッチングによって選択されるノード列は「定常なノード」だけである。同一性のチェックにおいて「見かけと機能に関する重要なアトリビュート（属性）が一致している場合に」同一のタグと判定する。これは、ページオーサが同一の見え方・機能をもったタグを細かい点でことなるアトリビュートを付加している可能性があるためである。本実施例の実装では以下のようなアトリビュートにより同一性を判定した。課題によっては、例えば、imgのsrcタグが完全にakamai等の付加分散システムによってコントロールされている場合、同一性判定からsrcタグははずすべきであろう。

基本："class", "id", "name", "style", "width", "height", "bgcolor"。

img系："alt", "src"。

link系："href"。

form系："action", "method", "type", "value"。

table系　"align", "valign", "rowspan", "colspan", "size", "color", "face"。,

上記において、「見かけに関するアトリビュート」とは"bgcolor"等、HTMLファイルの表示状態において見栄えに関するものである。「機能に関するアトリビュート」とは、link系の"href"やform系の"action"及び "method"等、HTMLファイルの表示状態には影響のないものである。

【００５３】

ステップ４： ターゲット・グループのツリー内の各ノードが過去ページに出現した頻度を「定常指数」として算出する。例えば、今、１２個の過去ページと比較を行い、あるエレメントがそのうち８ページに出現した場合、８／１２＝０．６７ が定常指数となる。この指数はこのような単純なパーセンテージだけでなく、頻度を示す数値であれば指数になり得る。

ステップ５： 定常ノードではないと判断されたノードを２回目の差分演算によ

り、「必須・更新ノード（必ず出現し、更新されるノード。「必須・更新ノード」は、本明細書において適宜、「更新ノード」と省略して呼ぶ。）」と「付加ノード（追加削除される可能性があり変動するノード）」に分類する。ステップ３では、テキスト・ノードに対して文字列が完全にマッチングした場合のみ同一であると定義した。このステップでは、文字列や画像がマッチしなくとも「テキスト・ノード（画像エレメント）が存在した場合」には同一であると判断する。またanchor(a)エレメントはhref属性が一致せずとも同一と判断する。 iframe等のsrc属性、href属性を持つものも同様に処理する。ステップ２のノード・リストに含まれずにこのステップにおけるノード・リストに含まれるノードは「必ず出現し、常に更新されるノード（テキスト、アンカー、画像）」であると言うことができる。

【００５４】

ステップ６： ステップ５でリストアップされたノードの頻度を算出する。この指数はステップ３と同様であり、単純なパーセンテージを使用することも可能である。

ステップ７：ステップ４と６の結果より各ノードを定常ノード、更新ノード、付加ノードに分類する。分類は指数を或る閾値で判定することで行う。例えば定常指数が７０％を超えたときに定常なノードであると判定する。ただし、手順1で算出したターゲット・サブツリーのうち、アノテーション・オーサによって指定されていないノード群、（ステップ１の対象ノード群をルート・ノードとして葉の方向に伸びるサブツリーには含まれないノード群）はすべて「pat:type属性」に「any」を設定。

【００５５】

このような差分演算の結果を図１３及び図１４に示す。図１３及び図１４はasahi.comのウェブ・コンテンツについての差分演算例を示したものであり、（ａ）はオリジナル（オリジナル・コンテンツ部分）、（ｂ）は差分結果をそれぞれ示している。（ｂ）において背景が色付きになっている部分が定常なノードであり、白い背景の部分が更新テキスト・ノードの部分である。図１３においては「全文 >>」という文字列が、図１４においては「最新ニュース」が定常なものと

して判別できていることが分かる。

【００５６】

ステップ８：さらに精度向上のために、イメージの種別判定を行う。これは、リストのbullet（ビュレット）や空白領域を確保するための「spacer GIF（スペーサGIF）」等を判定し、繰り返しパターンから除外するためである。spacer GIFは一つのページに複数使用され、かつ使用されるたびに指定サイズのことなるイメージとする。bullet イメージは一つのページに複数使用され、常に同じサイズで使用されるかサイズ指定のないイメージとする。次にパターン内のサブツリーの繰り返しを解析する。サブツリーの繰り返しパターンにいくつかの方法が存在するが、ここではシリアライズしたベクトルを対象に探索を行うことで比較的高速に検出を行うアルゴリズムを示す。

【００５７】

ステップ９：分類したツリー構造をシリアライズし、各ノードに関し以下の情報を算出して新たなベクトルを生成する。

距離ベクトル（ Distance vector）：次に出現する「同一レベル・同一タグタイプ・同一値ノード」のシリアライズされたベクトル上での距離。

例えば、以下のような例を考える。

ここで、更新ノードは「pat:type="updated"」と、付加ノードは「pat:type="inserted"」と表記した。

```
<div>
    <ul>
        <li><pat:text pat:type="updated"/></li>
        <li><pat:text pat:type="updated"/></li>
        <li><pat:text pat:type="updated"/>
            <img src="。。/new。gif" pat:type="inserted">
        </li>
    </ul>
    <ul>
        <li><pat:text pat:type="updated"/></li>
```

```
        <li><pat:text pat:type="updated"/></li>
        <li><pat:text pat:type="updated"/></li>
    </ul>
    <ul>
        <li><pat:text pat:type="updated"/></li>
        <li><pat:text pat:type="updated"/></li>
        <li><pat:text pat:type="updated"/></li>
    </ul>
</div>
```

【００５８】

図１５はDOMツリーの一例である。この例では、エレメントdiv、ul及びliに相当するノードは定常ノードであり、最下層のノードは更新テキスト・ノード又は追加イメージ・ノードとなっている。図１６は直列化されたノードのベクトルと各段の距離ベクトルとの関係を示している。図１６において（ａ）は直列化されたノードのベクトル（シリアライズ・ベクトル）を示し、（ｂ）～（ｅ）はそれぞれ１段、２段、３段及び４段の距離ベクトルを示す。なお、この直列化は深さ優先方式の直列化となっている。図１５のDOMツリーから図１６（ａ）への変換において、シリアライズ・ベクトルには「付加ノード(pat:type="inserted")」を組み込まない。これにより一時的に挿入されたコンテンツをパターンの算出から除外することができ、パターンのロバストネス（robustness）を高めることができる。例えば、図１８に示すようなパターンも図に示した部分を「付加ノード部分」として繰り返し判定から除外することができる。 付加ノードは後段の処理においてパターンに含められる。

【００５９】

また、ステップ８においてビュレット（bullet）イメージと判定された画像は異なった画像が用いられていても同一画像と判定する。これにより、例えば、図１９に示したようにbulletが変動する列挙パターンも繰り返しパターンとして検出可能になる。

【００６０】

さらに、「２つ目の同一ノードまでの距離」を示す「２段目の距離ベクトル（distance vector）」を算出する（図１６（ｄ））。同様に３段目（図１６（ｅ））、４段目（図１６（ｅ））と順次算出し、すべてのノードの値がベクトル長の１／３以上になるまで段数を増やす。これは最長繰り返しパターンの１回の繰り返し（iteration）がベクトル長の１／３以下であるからである。図の例ではベクトル長が２２ノードであるから４段目（図１６（ｆ））以降を算出する必要はない。

【００６１】

ステップ１０：ステップ７で算出したベクトルを基に繰り返しパターンを検出する。すなわち、距離ベクトルにおいて同一の距離が「２回繰り返し以上連続する部分」を探索する。例えば、距離「５」が連続した場合、それが１０以上連続した場合に繰り返しパターンとして検出する。これは、同じエレメントパターンが３回繰り返し以上連続していることを意味するからである。

【００６２】

図１７の例では、１段目と３段目にまたがってパターンを検出している。 このとき、２段目及び３段目の距離ベクトルに含まれていてもかまわない。ただし、このとき、繰り返しパターンがサブツリー間に「またがらない」ようにチェックする。例えば、下記のようなDOM構造があった場合に、6から10、11から15を繰り返しとして検出するのではなく、8から12、13から17を検出するようにチェックを行う。すなわち下位のノードの繰り返しの距離は、上位のノードの繰り返しをまたがって、検出されないようにする。

【００６３】

```
1:<ul>
2:    <li>
3:        <b>経済
4:        </b>
5:    </li>
6:    <li><pat:text pat:type="updated"/></li>
7:</ul>
```

```
8:<ul>
9:    <li><pat:text pat:type="updated"/></li>
10:    <li><pat:text pat:type="updated"/></li>
11:    <li><pat:text pat:type="updated"/></li>
12:</ul>
13:<ul>
14:    <li><pat:text pat:type="updated"/></li>
15:    <li><pat:text pat:type="updated"/></li>
16:    <li><pat:text pat:type="updated"/></li>
17:</ul>
```

【００６４】

ステップ１１：検出された繰り返し部分を<repeat>タグで囲み、繰り返しを除去する。繰り返し部分は同一の距離(図１７で「7」)が連続する部分に加えて、繰り返しの最後に対応する部分もパターンに加える。さらに、ステップ７においてシリアライズの際に除外したinsertedノードを対応する位置に挿入する。

【００６５】

```
<div>
   <repeat>
      <ul>
         <li><pat:text pat:type="updated"/></li>
         <li><pat:text pat:type="updated"/></li>
         <li><pat:text pat:type="updated"/>
            <img src="../new.gif" pat:type="inserted">
         </li>
      </ul>
   </repeat>
</div>
```

【００６６】

ステップ１２：分類したツリー構造をマッチング用のパターンとして整形する。

このアルゴリズムの出力例を示す。ただし便宜上、既存のパターン・マッチ記述ではなく、html記述にわずかにタグを追加するだけの独自表現を以下の説明では用いることにする。これは可読性を考慮したためであり、記述能力的には等価な既存言語に変換可能である（後述）。図２０及び 図２１はそれぞれ繰り返しを含むウェブ・コンテンツの例としてニュース・ライコス（News LYCOS）及びCNN.COMのウェブ・コンテンツのイメージを示している。また、図２２はtd内にtableが構造化・階層化コンテンツ用処理システム10個以上連続するウェブ・コンテンツのイメージを示している。これらウェブ・コンテンツから自動生成されたパターン(XML形式)を以下に示す。ベースタグセットはxhtmlであり、patネーム・スペースとしてパターンのためのタグが挿入されている。なお、図２１のウェブ・コンテンツでは、オーサは構造化・階層化コンテンツ用処理システム10個以上連続するtableの内の２個が選択されたとして、自動生成されたパターン(XML形式)を示す。

【００６７】

また、ここではネーム・スペースを利用して"pat"というプレフィックスで繰り返し等を表現する記法を用いたが、他のツリー正規表現記述に等価に置き換えることが可能であるものとする。例えばrelaxNGで利用されているTREXの記述力は本手法におけるパターンのために十分な記述力をもっており、本手法のパターン記述に使用することが可能である。これに関しては後述する。

【００６８】

図２０のウェブ・コンテンツから自動生成されたパターン(XML形式)

```
<table width="168">
 <tbody>
  <tr bgcolor="dedede">
   <td>
    <b>
     <span>トピックス</span>
    </b>
   </td>
```

```
  </tr>
  <pat:repeat>
    <tr bgcolor="ffffff">
     <td>
      <small>
       <a>
        <pat:text pat:type="updated">
       </a>
      </small>
     </td>
    </tr>
  </pat:repeat>
  <tr bgcolor="ffffff">
   <td>
    <small>
     <div align="right">
      <span>[</span>
      <a>
       <span>もっと見る</span>
      </a>
      <span>]</span>
     </div>
    </small>
   </td>
  </tr>
 </tbody>
</table>
```

【００６９】

図２１のウェブ・コンテンツから自動生成されたパターン(XML形式)

```
<table width="345">
 <tbody>
  <tr>
   <td bgcolor="#CC0000" style="background-color: #c00;">
    <span class="cnnMainHeaderBarText" style="color: #fff">
     <b>
      <span>?AMERICA AT HOME?</span>
     </b>
    </span>
   </td>
   <td bgcolor="#000033" style="background-color: #003;" width="60%" ali
gn="right">
    <span class="cnnMainHeaderBarText">
     <a style="color: #fff">
      <b>
       <span>more≫</span>
      </b>
     </a>
     <span>?</span>
    </span>
   </td>
  </tr>
  <tr>
   <td colspan="2">
    <div class="cnnMainT2List">
<!-- investigation -->
     <pat:repeat>
       <div style="padding-top: 3px; padding-bottom: 3px;">
        <li>
```

```
          <span class="cnnMainT2Area">
           <a>
            <pat:text pat:type="any">
           </a>
          </span>
         </li>
        </div>
       </pat:repeat>
       <div style="padding-top: 3px; padding-bottom: 3px;">
        <li>
         <span class="cnnMainT2Area">
          <a>
           <pat:text pat:type="any">
          </a>
         </span>
        </li>
       </div>
       <div style="padding-top: 3px; padding-bottom: 3px;">
        <li>
         <span class="cnnMainT2Area">
          <span>Fact Sheet: </span>
          <a>
           <pat:text pat:type="any">
          </a>
         </span>
        </li>
       </div>
  <!-- /investigation -->
     </div>
```

```
   </td>
  </tr>
 </tbody>
</table>
```

【００７０】

図２２のウェブ・コンテンツから自動生成されたパターン(XML形式)

```
<td width="99%">
 <pat:element pat:type="any">
 <table width="100%" pat:type="targetnode">
  <tbody>
   <tr bgcolor="dedede">
    <td>
     <b>
      <span>経済</span>
     </b>
     <small>
      <pat:text pat:type="any">
     </small>
    </td>
    <td align="right">
     <small>
      <a>
       <span>経済</span>
      </a>
      <span> | </span>
      <a>
       <span>企業</span>
      </a>
      <span> | </span>
```

```
      <a>
       <span>マーケット</span>
      </a>
     </small>
    </td>
   </tr>
  </tbody>
 </table>
<table width="100%"   pat:type="targetnode">
  <tbody>
   <tr>
    <td>
     <a>
      <b>
        <pat:text pat:type="any">
      </b>
     </a>
     <small>
      <nobr>
        <pat:text pat:type="any">
      </nobr>
     </small>
    </td>
   </tr>
   <tr>
    <td>
        <pat:text pat:type="any">
     <nobr>
        <pat:text pat:type="any">
```

```
      <a>
        <pat:text pat:type="any">
      </a>
        <pat:text pat:type="any">
    </nobr>
    <nobr>
        <pat:text pat:type="any">
      <a>
        <pat:text pat:type="any">
      </a>
        <pat:text pat:type="any">
    </nobr>
   </td>
  </tr>
 </tbody>
</table>
<pat:element pat:type="any">
</td>
```

【００７１】

タイプＢ：過去ページが存在しない場合のマッチング・パターンの算出

過去ページが存在しない場合とは、過去ページのキャッシングが行われていないときのみならず、日々生成されるURL等で頻繁に発生する。例えば新聞記事のURLのように日付がURLの一部として利用されている場合等明らかに過去のページは存在し得ない（http://www.asahi.com/international/update/1005/010.html）。また、検索結果のページ等クエリーの場合も同様である。このような場合、「隣接ページ」という概念を導入する。隣接ページとは、以下のような条件を持つページ群である。

【００７２】

（ａ） URLが近い。URLの近さはURLのエディットディスタンスにより定義する。

例：

ターゲット：http://www.asahi.com/international/update/1005/010.html

隣接URL：http://www.asahi.com/international/update/1005/012.html

（ｂ） レイアウトが近い。この判定にはテーブル構造の比較によりクラスタリング技術を利用する（例：前述した特許文献２）。この技術はテーブルの入れ子構造を基本として各ページのレイアウトをクラスタリングする手法であり、レイアウトの近いページのリストを得ることができる。

【００７３】

これらの条件に当てはまるページ群が「隣接ページ」である。以下、処理のステップを述べる。同様に、アノテーション・エディタを用いてアノテーション・オーサがすでにツリー上のあるノードをすでに指定した状態を考える。

ステップ１：隣接ページのリストを取得する。キャッシュサーバが隣接ページの算出アルゴリズムを持つものとし、キャッシュサーバから隣接ページのリストを取得する。各隣接ページは現在のもののみならず、過去の隣接ページも取得する。

ステップ２：各隣接ページと現在ターゲットになっているページの差分演算を行う。タイプＡのステップ２と同様に、差分演算のシリアライズを行う際に、テキスト・ノード、画像エレメントの同一性は「文字列もしくは画像が完全に同一」であることによって定義する。

ステップ３：ターゲット・グループのツリー内の各ノードが過去ページに出現した頻度を「定常指数」として算出する。

ステップ４：文字列や画像がマッチしなくとも「テキスト・ノード（画像エレメント）が存在した場合」には同一であると判断して各隣接ページとターゲット・ページの差分演算を行う。ステップ２のノード・リストに含まれずにこのステップにおけるノード・リストに含まれるノードは「必ず出現し、常に更新されるテキスト（画像）」であるといえることができる。

ステップ５：ステップ４でリストアップされたノードの頻度を算出する。この指数はステップ３と同様であり、単純なパーセンテージを使用することも可能である。

【００７４】

ステップ６：ステップ３と５との結果より各ノードを定常ノード、更新ノード、付加ノードに分類する。分類は指数を或る閾値で判定することで行う。例えば定常指数が７０％を超えたときに定常なノードであると判定する。この結果例を図２３～図２５に示す。図２３（ａ）及び（ｂ）はasahi.comのINDEXページのイメージと差分結果を対比して示している。図２４はasahi.comのスポーツ・ページのイメージを示し、図２５は図２４のイメージに基づく差分結果を示している。実際の差分演算は、多くの隣接ページに出現しているエリア程、青が濃くなるカラーで画面に表示されている。図２４では、固定的なインデックスリストの項目が定常となっており、図２３（ｂ）は実際のカラーイメージをモノクロイメージにして示しているために見え難いが、「天気」、「社会」、・・・、「今日の朝刊」の項目及び各項目の左側のボタンのエリアが青の濃い定常ノードとして検出されている。また、図２５において、記事本文はそのエリアの背景が白っぽい灰色で表示されており、記事本文は更新されるものとして検出されていることが分かる。

【００７５】

ここからはタイプＡのステップ８以降の処理と同様である。タイプＡとタイプＢの最も大きな違いは、比較するページの個数である。タイプＡでは確実な過去のページという比較対照が存在するため、数ページの比較で適切にノードの分類をすることができる。しかし、タイプＢでは隣接ページといういわば「確実ではない」及び「本質的に異なったレイアウトである可能性を含んだ」対象との差分演算を行わなければならない。そのため、できれば数百から数千ページのオーダーのページと差分演算を行った上で統計量として指数を算出することが望ましい。

【００７６】

次に、本発明により生成されたマッチング・パターンについて種々の利用態様を説明する。

○フリー・アノテーション：

フリー・アノテーションとは、XPathを持たず（もしくは大まかなポジションの

み）でページ内のどこにそのグループが出現してもマッチさせる手法である。図２６はフリー・アノテーションの概略説明図である。図２６において図１０と同一の要素は同符号で指示して、説明は省略する。ユーザ７５が所定のアクセサブルHTML９６の送信要求をトランスコーディング・モジュール７７へ出す。トランスコーディング・モジュール７７は、対応のウェブ・サーバ７８から対応のターゲットHTML７９を受け取り、該ターゲットHTML７９に関連付けられる全部のアノテーションをアノテーション・データベースに要求する。アノテーション・データベース及びアノテーション・セット９７における口は、それぞれ特定のグループを指すアノテーションに対応付けてマッチング・パターンを持っている。アノテーション・データベースはターゲットHTML７９の各サブツリーにマッチするマッチング・パターンを持ったアノテーション・セット９７を選択し、トランスコーディング・モジュール７７へ返す。トランスコーディング・モジュール７７は、アノテーション・データベースから返されたアノテーション・セット９７に基づいてターゲットHTML７９を変換（トランスコード）して作成したアクセサブルHTML９６をウェブ・クライアント７６へ送る。トランスコーディング・モジュール７７では、ターゲットHTML７９のトランスコードにおいて、ターゲットHTML７９におけるロバストな切り出し位置指定を実現できる。また、トランスコーディングに用いた場合、ページ内で移動するグループや、あるサイトのすべてのページに対してあるパターンにマッチするグループを検出するといった用途に応用できる。このフリー・アノテーション処理を従来からのダイナミック・マッチングの手法の後に行うことで漏れテキストやアノテーションのマッチしなかったページに対してアノテーションを付加できる可能性があり、フェイル・セーフなシステムを構築することができる。

【００７７】

図２７はすでに公知のダイナミック・マッチングと図２６のフリー・アノテーションとを組み合わせたフェイル・セーフ付きアノテーション処理についての概略説明図である。図２７において図１０及び図２６と対応する部分は同一の符号を付け、説明は省略する。トランスコーディング・モジュール７７は、第１段としてダイナミック・マッチングにおいてXPathについて全部のアノテーションが

ターゲットHTML７９にマッチするアノテーション・セットを探索する。もしあれば、そのアノテーションをトランスコーディング・モジュール７７へ送り、トランスコーディング・モジュール７７は、該アノテーション・セットに基づいてターゲットHTML７９をトランスコードして、トランスコーディッドHTML８１を作成し、トランスコーディッドHTML８１をウェブ・クライアント７６へ送る。もし、ダイナミック・マッチングにおいてマッチするアノテーション・セットがダイナミック・マッチング用アノテーション・データベース９９において探索できなければ、トランスコーディング・モジュール７７は、アノテーション・データベースにフリー・アノテーションの指示を出し、アノテーション・データベース８０からアノテーション・セット９７を受け取り、該アノテーション・セット９７に基づいてターゲットHTML７９をトランスコードして、トランスコーディッドHTML８１を作成し、トランスコーディッドHTML８１をウェブ・クライアント７６へ送る。

【００７８】

本手法には、ツリーの定常性を統計的手法を用いて算出しているために、「ページごとに大きくDOMツリー上で位置が変わる一連のノード群」をパターンとして算出することが困難であるという制限がある。例えば、あるテーブルが、リロードされるたびに、いかなる場所にも出現しうる場合、統計量として現れることは少ないと考えられる。そのため、本手法をもちいて検出可能な「フリーなグループ」とは「大きく変動しないデフォルトのポジションがある」ことが前提であり、その点で制限がある。ただし、アノテーションずれが発生するケースとしては「新たなtrが挿入されてずれる」「trの順序が入れ替わる」等の頻度が高いことが経験的に知られており、これらの変化に対して対応できる点で本手法は十分有効である。

【００７９】

[フリー・アノテーション利用例：アノテーション・エディタによるフリー・アノテーションの作成]

以下は、アノテーション・エディタにおけるオーサの操作手順である。

ステップ１：アノテーション・エディタで、任意の領域（DOMツリーのサブツリ

ー）を選択。

ステップ２： 新規グループ追加を指示。

ステップ３：グループ定義ダイアログにおいて「フリー・アノテーション」チェック・ボックスをチェック。これに伴い、 システムが自動的にマッチング・パターンを算出。

ステップ４： ユーザ（オーサ）は、アノテーション・エディタを用いてステップ３のマッチング・パターンについての他のページへの適用可能性を判断する。

【００８０】

［フリー・アノテーション利用例：フリー・アノテーション用サイト・パターン・アナライザによるアノテーションの修正］

フリー・アノテーションは、これまでのでサイト・パターン・アナライザに類似した管理アプリケーションが必要になる。図２８はフリー・アノテーション用サイト・パターン・アナライザ(SPA2)の画面予想図を示す。アノテーション・マッチングウィンドウの左側にはURLが並び、横軸にはフリー・アノテーションが並び、それぞれ各ページとのマッチングを表示している。アノテーションの番号をクリックすることでソートすることが可能である。オーサは、誤ってマッチしているパターンを発見した場合、以下のようなステップで修正を行う。

ステップ１：正しくマッチングしているURLを複数個選択。

ステップ２：誤ってマッチングしているURLを複数個選択。

この後、システムは、正しくマッチングしているURLにはすべてマッチし、誤っているグループにはマッチングしないようにマッチング・パターンを修正。

【００８１】

○従来のダイナミック・マッチングへの応用：

従来型のダイナミック・マッチング手法へは、XPathに付け加えるコンテンツ条件として本手法を用いることができる。図２９はダイナミック・マッチング手法にマッチング・パターンによるマッチングを組み込んだマッチング・システムの構成図である。図２９において、図２６の要素と同一のものは同符号で指示して、説明を省略する。アノテーション・データベース101では、ターゲットHTML79について、XPathによるマッチングに加えてマッチング・パターンによるマッチ

ングについても判定する。結果、判定精度が向上する。なお、アノテーション・データベース101の各アノテーション・セットにおいて、塗りつぶされた口はXPathに及びマッチング・パターンの両方にマッチしたアノテーションを意味する。

【００８２】

[従来のダイナミック・マッチングへの応用例：アノテーション・エディタによるグループに対する詳細条件としてのグループマッチングの追加]

オーサの操作手順は次の通りである。

ステップ１：アノテーション・エディタで、任意の領域（DOMツリーのサブツリー）を選択。これは標準的な操作と何ら代わるところはない。

ステップ２：新規グループ追加を指示。

ステップ３：オート- グループ定義ダイアログにおいて「詳細化」ボタンを押す。

これに伴い、システムが自動的にマッチング・パターンを算出する。標準的なPC（パーソナル・コンピュータ）で演算時間はタイプＡで数秒から数十秒、タイプＢで数十秒から数分の処理時間が必要になる。

ステップ４：オーサは、アノテーション・エディタを用いて他のページへの適用可能性を判断する。

【００８３】

[従来のダイナミック・マッチングへの応用例：サイト・パターン・アナライザによるダイナミック・マッチングアノテーションへの適用]

オーサの操作手順は次の通りである。

ステップ１：サイト・パターン・アナライザで誤ってマッチしているグループを探す。

ステップ２：- 正しくマッチしているページ及び誤ってマッチしているページを双方数ページずつを選択する。この操作はすでに実現されているXPathの半自動詳細化と同様である。

ステップ３： 一覧のうち、正しくマッチしているページから成るグループ群を選択し、「詳細化」を選択。

ステップ４：差分演算を用いて正しいグループ群が必ずマッチするマッチング・

パターンを自動生成。

ステップ５：生成されたマッチング・パターンが誤りグループとマッチしないことを確認。誤りグループとのマッチが発生してしまう場合は、従来からのXPathの半自動修正機能を用いてさらに条件を詳細化する。

【００８４】

次に隣接ページを用いた場合の精度について述べる。隣接ページをマッチング・パターンの生成に用いた場合、リストアップされる隣接ページによって生成されるマッチング・パターンが大きく左右されてしまう問題がある。図３０は或るウェブ・コンテンツの所定領域を隣接ページとの差分演算処理した結果を示している。（ａ）はマッチング・パターンを求めようとするターゲットウェブ・コンテンツ、（ｂ）は差分演算によりによりノードの種類を検出した結果を示す。（ｂ）において、「関連情報」の領域の背景は、適宜変更される見出し文の領域の背景と同じく、薄い色になっている。「関連情報」という文字列は明らかに定常であり、マッチング・パターンに組み入れられるべきである。しかし、隣接ページとの差分を行った場合、このような場所の変動と大きい文字列を定常であると判断することは難しい。この問題には本発明では２つの方法で対処する。

（ａ）隣接ページの厳選。前述したレイアウトのクラスタリング技術をつかい同一のレイアウトを用いていると考えられるページのみをリストアップする。

（ｂ） 誤り修正のためのインタフェース。前述のサイト・パターン・アナライザSPA，SPA2はこのような誤りを修正するためのインタフェースを持つ。

【００８５】

○切り出した情報種別の判定：

図３１は株価情報のウェブ・コンテンツから株価の数値を切り出すことについてのマッチング・パターンの利用説明図である。（ａ）は株価情報を提示するウェブ・コンテンツを示し、（ｂ）過去ページとの差分演算による検出した定常ノードを示している。株価情報のテーブル等からの株価の数値の切り出しはアノテーションのマッチング・パターンとして組み入れることも考えられる。例えば「12時13分更新」というテキストに対して、

<pat:data pat:type="date" pat:format="HH時MM分更新" pat:xpath="table[1]/

tr[1]/td[3]/text()[1]"/>

といった記述でHH，MMという時間情報を切り出すことが可能である。このように数値データ、テキストデータの切り出しをマッチンパターンに組み入れることも可能である。これにより、RSSやWSXLあるいはVoiceXMLへの変換に大きな効果があると考えられる。

【００８６】

○ダイナミック・アノテーションの手法との融合・XPathセットマッチ高速化アルゴリズムの利用：

今回のサブツリーのマッチングをXPathセットのマッチングと捉えることも可能である。こうすることによりこれまでに提案しているXPathセットの高速マッチングの手法を利用することが可能である。ただし、repeatを用いた繰り返しやpat:type="inserted"を表現することができないため、すべてのマッチング・パターンを変換することはできない。

【００８７】

(グループのXPathをルートとして)

/tr[1]

/tr[1]/td[1][@bgcolor="#006699"]

/tr[1]/td[1][@bgcolor="#006699"]/font[1][@color="#ffffff"]

/tr[1]/td[1][@bgcolor="#006699"]/font[1][@color="#ffffff"]/text()[1]

/tr[1]/td[1][@bgcolor="#006699"]/font[1][@color="#ffffff"]/b[1]

/tr[2]

/tr[2]/td[1]

/tr[2]/td[1]/small[1]

/tr[2]/td[1]/small[1]/li[1]

/tr[2]/td[1]/small[1]/li[1]/a[1]

/tr[2]/td[1]/small[1]/li[1]/a[1]/text()[1]

/tr[2]/td[1]/small[1]/li[2]

/tr[2]/td[1]/small[1]/li[2]/a[1]

/tr[2]/td[1]/small[1]/li[2]/a[1]/text()[1]

・・・

/tr[2]/td[1]/small[1]/li[6]

/tr[2]/td[1]/small[1]/li[6]/a[1]

/tr[2]/td[1]/small[1]/li[6]/a[1]/text()[1]

/tr[2]/td[1]/small[1]/li[6]/div[1][@align="right"]

/tr[2]/td[1]/small[1]/li[6]/div[1][@align="right"]/text()[1]

/tr[2]/td[1]/small[1]/li[6]/div[1][@align="right"]/a[1]

/tr[2]/td[1]/small[1]/li[6]/div[1][@align="right"]/a[1]/text()[1]

/tr[2]/td[1]/small[1]/li[6]/div[1][@align="right"]/text()[2]

【００８８】

また、従来型ダイナミック・アノテーション・マッチングの手法と組み合わせる場合、他のグループが持っているXPathとマッチング・パターンから生成されるXPathをすべて列挙することにより、一体的に取り扱うことも可能である。

【００８９】

○p，br，bタグ及びテキスト・ノードがランダムに出現する場合の対応：

或るコンテンツの本文等に、p，br，bタグ及びテキスト・ノードがランダムに出現する場合がある。このような場合に対処するために、p，br，bタグに増減があってもマッチさせることのできるマッチング・パターンを生成する必要がある。そのため、p，br，bタグの連続がターゲット・ページ、過去ページに出現した場合は、すべてを「ANY」ノードにするという処理を行う。すなわち、「ANY」マッチング・パターンにおける正規表現として利用する。

【００９０】

○既存のツリーマッチング記述言語形式の生成：

今回、説明において独自のパターン記述を用いたが、これは等価なパターン・マッチング記述言語に変換可能である。しかし、元のツリー構造を保存できない、アトリビュートの厳密な記述が必要になるという点で煩雑になり可読性が低下するため説明には用いなかった。そこで、今回用いた記法を既存のパターン・マッチ言語(relaxNG形式)に変換する手法の一部を紹介する。

まず以下のようなパターンを考える。

【００９１】

```
<table width="168">
 <tbody>
  <pat:repeat>
    <tr bgcolor="ffffff">
     <td>
      <small>
       <a>
        <pat:text pat:type="any">
       </a>
      </small>
     </td>
    </tr>
  </pat:repeat>
  </tbody>
</table>
```

【００９２】

これをrelaxNG形式に変換例を以下に示す。ただし、アトリビュートの記述は一部省略した。relaxNGはXML文書全体のSchemaを記述するように設計されているため、本来、ルートタグを含めてすべてにマッチするパターンを記述するようにできている。ここではその枠組みをサブツリーのマッチングに用いる。そのため、実装としては、次のような２個のステップで処理を行うことになる。

ステップ１：HTML内のtableタグをすべてリストアップ

ステップ２：tableを一つずつ、マッチング・パターンとマッチするか評価

以下のサンプルはそのような実装を前提としている。なお、以下は、relaxNG形式による記述例である。

【００９３】

```
<?xml version="1.0" ?>
<grammar xmlns="http://relaxng.org/ns/structure/0.9">
```

```
<start>
  <element name="table">
    <attribute name="width">
      <value>168</value>
    </attribute>
    <zeroOrMore>
      <choice>
        <ref name="freeAttributesTABLE"/>
      </choice>
    </zeroOrMore>
    <element name="tbody">
    <zeroOrMore>
      <choice>
        <ref name="freeAttributesTBODY"/>
      </choice>
    </zeroOrMore>
  <oneOrMore>
        <element name="tr">
          <attribute name="bgcolor">
            <value>ffffff</value>
          </attribute>
          <zeroOrMore>
            <choice>
              <ref name="freeAttributeTR"/>
            </choice>
          </zeroOrMore>
          <element name="td">
            <zeroOrMore>
              <choice>
```

```
            <ref name="freeAttributesTD"/>
          </choice>
        </zeroOrMore>
        <element name="small">
          <zeroOrMore>
            <choice>
              <ref name="freeAttributesSMALL"/>
            </choice>
          </zeroOrMore>
          <element name="a">
            <zeroOrMore>
              <choice>
                <ref name="freeAttributesA"/>
              </choice>
            </zeroOrMore>
            <text/>
           </element>
         </element>
      </element>
    </element>
  </oneOrMore>
  </element>
</element>
</start>

<define name="freeAttributesTD">
<attribute>
  <anyName>
    <except>
```

<name>width</name>
</except>
<except>
<name>height</name>
</except>
省略。TDタグにおいてここにはマッチングにおいて無視できないアトリビュートの列を記述する。
</anyName>
</attribute>
省略。以下、freeAttributes定義が各タグごとに並ぶ。
</grammer>

【００９４】

○マッチング・パターン生成能力の点から見た本手法の制限：
ツリーの正規表現の自由度としてはrepeat(繰り返し)とembed(入れ子)の２種類が存在することが知られている。このうち、本手法はrepeatのみ検出することができる。これは、HTMLの領域マッチングに用いるという性質上、入れ子構造による規則性を記述する必要性が非常に低いことに基づいている。そのため、統計情報を用いるという基本アイディアに基づいて入れ子構造を算出するアルゴリズムに拡張することも可能である。

【００９５】

［その他の実施例１：アノテーションによるトランスコーディング］
従来のアノテーションシステムによって、アノテーションがマッチせず漏れたページ、漏れた情報等に対して今回のフリー・アノテーションを用いてカバーする「フェイル・セーフ」システムを構築可能である。これはトランスコーディングの品質保証を通じてビジネスに大きく貢献する。さらに、本発明によりマッチング条件の詳細化を行うことで、アノテーション修正の手間が減り、アノテーション・オーサリング時間を短縮することができる。これもビジネスに大きく貢献する機能である。さらに、従来のトランスコーディングにおいてXpathの文字列マッチングを用いてしか判定することのできなかったグループ部分をフリー・アノ

テーションよってカバーすることができる。図３２は所定の定常ノードが移動するウェブ・コンテンツの例を示している。図３２のなかで「LYCOSサービス」や「関連トピックス」といった部分は、上下に場所が移動するばあいがあり、従来の枠組みでは取り扱いにくかった。このようなグループに対しても本手法であれば対処することが可能である。

【００９６】

［その他の実施例２：リンク・リストの切り出しによりRSSの生成］

RSSとはRich Site Summaryと呼ばれ、XML形式であるサイトの要約を定義し、提供することで、サイトのサマリーをさまざまに再利用できるようにする規格である。従来はサイトごとに、CGI等をもちいてダイナミックにこのRSSを生成していた。しかし本発明を用いることにより、ウェブ・ページからダイナミックに生成可能になる。まず、アノテーション・エディタを用いてサイトのトップ・ニュースのリストにあたるリンク・リストを指定する、フリー・アノテーションを作成する。このグループに対し「RSSアトリビュート」を付加する。RSSエンジンは、このフリー・アノテーションを用いてウェブ・ページから直接RSS形式のデータを生成する。このような「特定の部分のみを指定するグループ」はXPathマッチングを用いた従来のアノテーションでは困難である。例えば、前述した図２０に係るパターン(XML形式)の例では、<pat:text pat:type="any">に示した部分はその日のトップ記事の各タイトルになっている。そのため、パターン・マッチの過程でワイルドカード部分を切り出すことで以下のようなRSS記述を自動生成することが可能になる。

【００９７】

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
  xmlns="http://purl.org/rss/1.0/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xml:lang="ja">
  <channel rdf:about="http://news.lycos.co.jp/topics/rss.rdf">
    <title>News LYCOS 最新Topics</title>
```

```
    <link>http://news.lycos.co.jp/topics</link>
    <items>
      <rdf:Seq>
        <rdf:li rdf:resource="http://news.lycos.co.jp/topics"/>
      </rdf:Seq>
    </items>
  </channel>
  <item rdf:about="http://news.lycos.co.jp/topics/society/maff.html">
    <title>諫早湾 緊迫の中、工事再開</title>
    <link>http://news.lycos.co.jp/topics/society/maff.html</link>
  </item>
  <item rdf:about="http://news.lycos.co.jp/topics/world/operation.html">
    <title>アルカイダ戦闘員 米に投降</title>
    <link>http://news.lycos.co.jp/topics/world/operation.html</link>
  </item>
  <item rdf:about="http://news.lycos.co.jp/topics/computer/ms.html">
    <title>マイクロソフト家電分野進出</title>
    <link>http://news.lycos.co.jp/topics/computer/ms.html</link>
  </item>
・・・以下省略
</rdf:RDF>
```

【００９８】

[その他の実施例３：Webページの部分切り出しによるWEB サービス化]

Web サービスはXMLの交換により、様様なサービス、アプリケーションを提供する技術であるが、本発明を用いることにより、すでに存在するウェブ・ページのトランザクションの一部を切り出す形で容易に提供することができる可能性がある。図３３は部分切り出しに利用する利用元のウェブ・コンテンツを例を示している。このページはあるニュース・サイト（ZDNET）内の過去の記事をキーワード検索し、提示するページである。このページをベースにしてキーワード検索を

行うWebサービスを構築することができる。指定する必要のあるグループは２つ。一つは検索のためのフォーム部分１０３（図３３）である。この領域は不動部分から構成されており、マッチング・パターンは生成し易い。

【００９９】

次はフォーム部分１０３をHTMLで表現したものである。

```
<select name="idxname" size="1" tabindex="2">
    <option value="" selected>All ZDNet
    <option value="news">ZDNN
    <option value="zdii">ZDII
. . . . . .
</select>
```

【０１００】

次は、上記のHTMLから自動生成したＳｃｈｅｍａの一部（ＲｅｌａｘＮＧ形式）である。このＳｃｈｅｍａは図３４のＳｃｈｅｍａ（１）として利用される。

```
<element name= “idxname” >
<choice>
  <element name="option">
  <element name="value">
    <string></string>
  </element>
  </element>
  <element name="option">
  <element name="value">
    <string>news</string>
  </element>
  </element>
  <element name="option">
  <element name="value">
    <string>zdii</string>
```

```
  </element>
  </element>
</choice>
</element>
・・・・・・・
```

【０１０１】

図３４は図３３のウェブ・コンテンツからWeb サービスを自動生成する処理過程を示す。切り出されたフォームに対して、以下のような入力のためのXML Schema（図３４のスキーマ（Schema）(1)とこのXMLを元のHTMLフォームに変換するためのXSLT （図３４のXSLT(2)）を自動生成することが可能である。

【０１０２】

```
<web_form_based_service action="./index.cgi" method="GET">
    <text>検索キーワード</text>
    <idxname>検索範囲指定</idxname>
    <max>最大検索結果数</max>
</web_form_based_service>
```

【０１０３】

さらにボキャブラリーの変更、自動生成のXSLT・XML Schema・WSDLの修正を行う必要があるが、プロトタイピングを行い、詳細な開発のベースとしての利用は可能であろう。

このようにWEBフォームを用いると完全ではないものの、比較的容易にWEBサービスのプロトタイピングを行うことができる。これはこれまでもCHIP等の技術でも一部実現可能である。

【０１０４】

問題は、検索結果の部分１０４（図３３）である。検索結果の部分１０４は変動するコンテンツがダイナミックに生成される部分であり、パターン化が非常に困難である。しかし、本発明を用いることにより、定常ノード、更新ノード、付加ノードを判別し、さらには繰り返しパターンを検出することができるため、以下のようなパターンを自動的に生成することができる（図３４のパターンによる

切り出し(5))。次のパターン記述は、図３４のパターンによる切り出し(5)に相当するものであり、RelaxNGではなく、独自形式である。

【０１０５】

```
<h2>検索結果</h2>
    <p>
        参考ヒット数:  [
        <pat:text pat:type="any"/>
        <pat:text pat:type="any" pat:format="[0-9]+"/>
        ]
    </p>
    <p>
        <strong>
            検索式にマッチする
            <pat:text pat:type="any"  pat:format="[0-9]+"/>
            個の文書が見つかりました。
        </strong>
    </p>
<dl>
<repeat>
    <a>
        <pat:text pat:type="any"/>
        <b>
            <font color=blue>
                <pat:text pat:type="any"/>
            </font>
        </b>
        <pat:text pat:type="any"/>
        <pat:text pat:type="inserted"/>
    </a>
```

```
    <font color=red size=-2>
        (
        <em>
            <pat:text pat:type="any"/>
        </em>
        )
    </font>
    <br>
      <pat:text pat:type="any"/>
      <b>
        <font color=blue>
            <pat:text pat:type="any"/>
        </font>
    </b>
    <pat:text pat:type="any"/>
      <pat:text pat:type="inserted"/>
      <br>
    <font color=green>
        <pat:text pat:type="any"/>
    </font>
    <br><br>
</repeat>
```

【０１０６】

このパターンから、結果部分の切り出しを行い、出力のXMLをここから生成することができる。そして、繰り返し部分はitemize、繰り返しから外れる部分の更新部分を特別なタグで出力するXML Schema（図３４（4））及び切り出された部分HTMLをXML形式に変換するためのXSLT（図３４（3））、XMLをHTMLに復号するためのXSLT（図３４（6））を自動生成することが可能である。

【０１０７】

［その他の実施例４：情報統合（Information Aggregator）への適用：
ウェブ・ページの一部分を切り出して、情報を統合することは、IBM PortalServer等のポータル構築システムや、IBM mySiteOutliner等の情報抽出・提示システムにおいて広く行われている。本発明はこれらのシステムに応用することが可能である。例えば、IBM mySiteOutlinerでは、ウェブ・ページから、ヘッドライン・リンク・リストを抜き出すために以下のようなXPathを定義ファイルの中に保持している。

【０１０８】

```
<ClippingDefinition>
    <id>2</id>
    <links>
    <link title="Club IBM トップ・ページ">http://www.ibm.com/jp/pc/clubibm/index.html</link>
    </links>
    <urldata>
        <url source="Club IBM">http://www.ibm.com/jp/pc/clubibm/index.html</url>
        <xpathlists>
          <xpath name="本文">
/html[1]/body[1]/table[2]/tbody[1]/tr[1]/td[2]/table[2]/tbody[1]/tr[5]/td[2]/table[1]/tbody[1]/tr[1]/td[1]/table[2]/tbody[1]/tr[2]/td[1]
          </xpath>
        </xpathlists>
    </urldata>
</ClippingDefinition>
```

【０１０９】

切り出し部位の指定は下線部のXPathに依存している。通常このようなXPath形式はレイアウトの変更に対して弱く、メンテナンスに大きな負担がかかるという問題がある。つまり、レイアウトの変更を人間が監視し、変更があった場合には

人手で再度正しいXPathをauthoringする必要がある。mySiteOutlinerの場合は、対象としているのが自社内ページコンテンツの切り出してあるため、レイアウト変更が事前にわかるため、変更されると同時に修正されたXMLファイルをユーザに配信することでこの問題に対処している。しかし管理コストの問題が依然として存在している。

【０１１０】

これに対し、本発明を適用することによって以下のように、マッチング・パターンを自動生成可能である。このパターンは対象としているテーブルのコンテンツ、とくに定常的な「新着情報」といった文字列や、テーブルのアトリビュートを手がかりとしているため、これらに対する変更が発生しない限りずれることはない。現状でずれが発生してしまうbody直下へのtableの挿入、上位テーブルタグにおけるtrの挿入、視覚的には影響がないdivタグspanタグの上位ノードへの挿入に対して完全にロバストである点で優れている。

```
<tr>
    <td width="440" height="20" bgcolor="#CCCCCC"> 新着情報</td>
</tr>
<tr>
    <td>
        <table border="0" cellpadding="0" cellspacing="2">
            <tbody>
                <repeat>
                    <tr>
                        <td>
                            <pat:img pat:img_type="bullet"/>
                        </td>
                        <td>
                            <a>
                      <font color="#006699"><pat:text pat:type="any"/></
font>
```

```
                    </a>
                </td>
            </tr>
        </repeat>
    </tbody>
</table>
</td>
</tr>
```

【０１１１】

このパターンにおいてロバストネスが損なわれる場合として例えば、以下のようなケースが考えられる。

（ａ）同一のパターンがマッチするコンテンツが同一ページ上に挿入される。

（ｂ）背景色、フォントカラー等アトリビュートの変更

（ａ）に関しては、視覚的にも同一の領域が出現することを意味しており、稀なケースであると考えられる。2に関しては再度パターンを生成するほか対処方法がない。しかし本発明では、レイアウト変更前のページをも統計量算出に用いることで両方のレイアウトに対してロバストなパターンを生成できる可能性がある点で2の問題に対しても対処可能である。

【０１１２】

まとめとして本発明の構成に関して以下の事項を開示する。

（１）：ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う構造化・階層化コンテンツ用処理装置であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定手段、

前記ターゲット・コンテンツに対する過去の複数個の構造化・階層化コンテン

ツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出手段、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成手段、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類手段、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成手段、

を有していることを特徴とする構造化・階層化コンテンツ用処理装置。

（２）：前記所定の処理とは、該構造化・階層化コンテンツのコンテンツ部分への関連情報の関連付けであることを特徴とする（１）記載の構造化・階層化コンテンツ用処理装置。

（３）：前記関連情報はアノテーションを含むことを特徴とする（２）記載の構造化・階層化コンテンツ用処理装置。

（４）：前記所定の処理とは、構造化・階層化コンテンツのコンテンツ部分を他の構造化・階層化コンテンツに利用するために該構造化・階層化コンテンツの該コンテンツ部分をコピーする処理であることを特徴とする（１）記載の構造化・階層化コンテンツ用処理装置。

（５）：構造化・階層化コンテンツとはウェブ・コンテンツであることを特徴とする（１）～（４）のいずれかに記載の構造化・階層化コンテンツ用処理装置。

【０１１３】

（６）：ターゲット・サブツリーのノードを、定常ノード、更新ノード及び付加ノードに分類する前記分類手段を有していることを特徴とする（１）～（５）のいずれかに記載の構造化・階層化コンテンツ用処理装置。

（７）：検出する前記出現態様として、（Ｎ１）被検出ノードがターゲット・コンテンツ部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に同一となって出現態様、及び（Ｎ２）被検出ノードがターゲット・コンテンツ

部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に異なっている出現態様を含む前記出現態様検出手段、及び

統計情報により（Ｎ１）の出現態様による出現頻度が第１の閾値以上であると判明したノードは定常ノードに分類し、統計情報により（Ｎ２）の出現態様による出現頻度が第２の閾値以上であると判明したノードは更新ノードに分類し、定常ノード及び更新ノード以外のノードは付加ノードに分類する前記分類手段、

を有していることを特徴とする（６）記載の構造化・階層化コンテンツ用処理装置。

（８）：前記マッチング・パターン生成手段は、

定常ノード、更新ノード及び付加ノードの分類に基づいてターゲット・サブツリーにおける繰り返し部分を検出する繰り返し部分検出手段、及び

該繰り返し部分の存在情報を含む前記マッチング・パターンを生成する繰り返し情報付きマッチング・パターン生成手段、

を有していることを特徴とする（６）又は（７）記載の構造化・階層化コンテンツ用処理装置。

（９）：前記分類手段は、

イメージに係るノードについて、該ノードが空白領域を確保するためのスペーサ用イメージに係るノードであるか否かを検出するスペーサ用イメージ検出手段、

イメージに係るノードについて、該ノードが繰り返して同一サイズで複数個使用されるビュレット・イメージに係るノードであるか否かを検出するビュレット・イメージ検出手段、

スペーサ用イメージに係るノードは付加ノードと分類する第１の分類付け手段、

ビュレット・イメージに係るノード同士は、その表示内容が異なっても定常ノード、更新ノード又は付加ノードの同一分類に割り当てる第２の分類付け手段、

を有していることを特徴とする（８）記載の構造化・階層化コンテンツ用処理装置。

（１０）：ターゲット・コンテンツに対する過去の構造化・階層化コンテンツが

存在しない場合には、過去の各構造化・階層化コンテンツの代わりに該ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択しターゲット・コンテンツに係るターゲット・サブツリーと各隣接構造化・階層化コンテンツに係るツリーと対照する前記対照手段、
を有していることを特徴とする（１）～（９）のいずれかに記載の構造化・階層化コンテンツ用処理装置。

【０１１４】

（１１）：ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う構造化・階層化コンテンツ用処理装置であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定手段、

前記ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと各隣接構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出手段、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成手段、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類手段、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成手段、
を有していることを特徴とする構造化・階層化コンテンツ用処理装置。

（１２）：ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判

定が正であれば該構造化・階層化コンテンツについて所定の処理を行う構造化・階層化コンテンツ用処理方法であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定ステップ、

前記ターゲット・コンテンツに対する過去の複数個の構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出ステップ、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成ステップ、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類ステップ、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成ステップ、

を有していることを特徴とする構造化・階層化コンテンツ用処理方法。

（１３）：前記所定の処理とは、該構造化・階層化コンテンツのコンテンツ部分への関連情報の関連付けであることを特徴とする（１２）記載の構造化・階層化コンテンツ用処理方法。

（１４）：前記関連情報はアノテーションを含むことを特徴とする（１３）記載の構造化・階層化コンテンツ用処理方法。

（１５）：前記所定の処理とは、構造化・階層化コンテンツのコンテンツ部分を他の構造化・階層化コンテンツに利用するために該構造化・階層化コンテンツの該コンテンツ部分をコピーする処理であることを特徴とする（１２）記載の構造化・階層化コンテンツ用処理方法。

【０１１５】

（１６）：構造化・階層化コンテンツとはウェブ・コンテンツであることを特徴とする（１２）～（１５）のいずれかに記載の構造化・階層化コンテンツ用処理方法。

（１７）：ターゲット・サブツリーのノードを、定常ノード、更新ノード及び付加ノードに分類する前記分類ステップを有していることを特徴とする（１２）～（１６）のいずれかに記載の構造化・階層化コンテンツ用処理方法。

（１８）：検出する前記出現態様として、（Ｎ１）被検出ノードがターゲット・コンテンツ部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に同一となって出現態様、及び（Ｎ２）被検出ノードがターゲット・コンテンツ部分及び対照構造化・階層化コンテンツの両方に出現しその内容が相互に異なっている出現態様を含む前記出現態様検出ステップ、及び

統計情報により（Ｎ１）の出現態様による出現頻度が第１の閾値以上であると判明したノードは定常ノードに分類し、統計情報により（Ｎ２）の出現態様による出現頻度が第２の閾値以上であると判明したノードは更新ノードに分類し、定常ノード及び更新ノード以外のノードは付加ノードに分類する前記分類ステップ、

を有していることを特徴とする（１７）記載の構造化・階層化コンテンツ用処理方法。

（１９）：前記マッチング・パターン生成ステップは、

定常ノード、更新ノード及び付加ノードの分類に基づいてターゲット・サブツリーにおける繰り返し部分を検出する繰り返し部分検出ステップ、及び

該繰り返し部分の存在情報を含む前記マッチング・パターンを生成する繰り返し情報付きマッチング・パターン生成ステップ、

を有していることを特徴とする（１７）又は（１８）記載の構造化・階層化コンテンツ用処理方法。

（２０）：前記分類ステップは、

イメージに係るノードについて、該ノードが空白領域を確保するためのスペーサ用イメージに係るノードであるか否かを検出するスペーサ用イメージ検出ステップ、

イメージに係るノードについて、該ノードが繰り返して同一サイズで複数個使用されるビュレット・イメージに係るノードであるか否かを検出するビュレット・イメージ検出ステップ、

スペーサ用イメージに係るノードは付加ノードと分類する第１の分類付けステップ、

ビュレット・イメージに係るノード同士は、その表示内容が異なっても定常ノード、更新ノード又は付加ノードの同一分類に割り当てる第２の分類付けステップ、

を有していることを特徴とする（１９）記載の構造化・階層化コンテンツ用処理方法。

【０１１６】

（２１）：ターゲット・コンテンツに対する過去の構造化・階層化コンテンツが存在しない場合には、過去の各構造化・階層化コンテンツの代わりに該ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択しターゲット・コンテンツに係るターゲット・サブツリーと各隣接構造化・階層化コンテンツに係るツリーと対照する前記対照ステップ、

を有していることを特徴とする（１２）～（２０）のいずれかに記載の構造化・階層化コンテンツ用処理方法。

（２２）：ネットワークを介して配信される構造化・階層化コンテンツが所定のマッチング・パターンとマッチするコンテンツ部分を含むか否かを判定し、該判定が正であれば該構造化・階層化コンテンツについて所定の処理を行う構造化・階層化コンテンツ用処理方法であって、

マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（以下、該構造化・階層化コンテンツを「ターゲット・コンテンツ」と言う。）におけるマッチング・パターンの抽出部分としてのターゲット・コンテンツ部分を含む範囲に係るターゲット・サブツリーを設定するターゲット・サブツリー設定ステップ、

前記ターゲット・コンテンツに対する複数個の隣接構造化・階層化コンテンツを選択し前記ターゲット・コンテンツに係るターゲット・サブツリーと各隣接構

造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する出現態様検出ステップ、

過去の複数個の構造化・階層化コンテンツに基づいて該ターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する統計情報生成ステップ、

前記出現態様検出結果及び前記統計情報に基づいてターゲット・サブツリーの各ノードを分類する分類ステップ、及び

該分類に基づいて前記ターゲット・コンテンツ部分についてのマッチング・パターンを生成するマッチング・パターン生成ステップ、

を有していることを特徴とする構造化・階層化コンテンツ用処理方法。

（２３）：（１２）～（２２）のいずれかに記載の構造化・階層化コンテンツ

【０１１７】

【発明の効果】

本発明では、一部切り出し及び共通のアノテーションの使い回し等の処理対象としての構造化・階層化コンテンツであるか否かを同定する（identify）ために、XPathではなく、マッチング・パターンを使用する。結果、同定対象としての構造化・階層化コンテンツにおいて、同定コンテンツ部分が適宜、移動する場合にも、柔軟に対処できる。

【０１１８】

本発明では、ターゲット・コンテンツに対する過去及び／又は隣接の構造化・階層化コンテンツを調べ、ターゲット・サブツリーにおける各ノードについての出現態様及び該出現態様の出現頻度に係る統計情報に基づいて各ノードを分類して、マッチング・パターンを生成する。結果、構造化・階層化コンテンツであるか否かを同定するために、有意義なマッチング・パターンを生成することがふできる。

【図面の簡単な説明】

【図１】

ウェブ・コンテンツ処理装置１４を装備する構造化・階層化コンテンツ用処理システム１０の構成図ある。

【図２】

構造化・階層化コンテンツ用処理装置１８のブロック図である。

【図３】

マッチング・パターン生成手段３０のより具体的なブロック図である。

【図４】

分類手段２９のより具体的なブロック図である。

【図５】

過去の構造化・階層化コンテンツに基づいてマッチング・パターンを生成する方法のフローチャートである。

【図６】

図５のマッチング・パターン生成方法において生成されたマッチング・パターンを使用するマッチング判定方法のフローチャートである。

【図７】

図５のマッチング・パターン生成ステップ（Ｓ５１）をより具体的に示すフローチャート部分である。

【図８】

分類手段２９のより具体的なブロック図である。

【図９】

ターゲット・コンテンツに対して隣接する複数個の構造化・階層化コンテンツに基づいてマッチング・パターンを生成する方法のフローチャートである。

【図１０】

ウェブ・コンテンツ用処理装置７４の構成図である。

【図１１】

DPマッチングの概略説明図である。

【図１２】

差分演算にDPマッチングを適用した概略説明図である。

【図１３】

sahi.comのウェブ・コンテンツについての第１の差分演算例を示す図である。

【図１４】

asahi.comのウェブ・コンテンツについての第２の差分演算例を示す図である。

【図１５】

DOMツリーの一例である。

【図１６】

直列化されたノードのベクトルと各段の距離ベクトルとの関係を示す図である。

【図１７】

各段の距離ベクトルを対比して示す図である。

【図１８】

繰り返し部の端部に付加ノード部分をもつウェブ・コンテンツを示す図である。

【図１９】

ビュレットが変動する列挙パターンももつウェブ・コンテンツを示す図である。

【図２０】

繰り返しを含むウェブ・コンテンツの例としてニュース・ライコス（News LYCOS）のイメージを示す図である。

【図２１】

繰り返しを含むウェブ・コンテンツの例としてCNN.COMのウェブ・コンテンツのイメージを示す図である。

【図２２】

td内にtableが構造化・階層化コンテンツ用処理システム10個以上連続するウェブ・コンテンツのイメージを示す図である。

【図２３】

asahi.comのINDEXページのイメージと差分結果を対比して示す図である。

【図２４】

asahi.comのスポーツ・ページのイメージを示す図である。

【図２５】

図２４のイメージに基づく差分結果を示す図である。

【図２６】

フリー・アノテーションの概略説明図である。

【図２７】

すでに公知のダイナミック・マッチングと図２６のフリー・アノテーションとを組み合わせたフェイル・セーフ付きアノテーション処理についての概略説明図である。

【図２８】

フリー・アノテーション用サイト・パターン・アナライザ(SPA2)の画面予想図を示す図である。

【図２９】

ダイナミック・マッチング手法にマッチング・パターンによるマッチングを組み込んだマッチング・システムの構成図である。

【図３０】

或るウェブ・コンテンツの所定領域を隣接ページとの差分演算処理した結果を示す図である。

【図３１】

株価情報のウェブ・コンテンツから株価の数値を切り出すことについてのマッチング・パターンの利用説明図である。

【図３２】

所定の定常ノードが移動するウェブ・コンテンツの例を示す図である。

【図３３】

部分切り出しに利用する利用元のウェブ・コンテンツを例を示す図である。

【図３４】

図３３のウェブ・コンテンツからWeb サービスを自動生成する処理過程を示す図である。

【符号の説明】

１０　構造化・階層化コンテンツ用処理システム

１２　インターネット

１４　ウェブ・コンテンツ処理装置

１５　ウェブ・クライアント

１６　ウェブ・サーバ

１８　構造化・階層化コンテンツ用処理装置

２０　ターゲット・コンテンツ

２１　ターゲット・コンテンツ部分

２５　ターゲット設定手段

２６　構造化・階層化コンテンツ・データベース

２７　出現態様検出手段

２８　統計情報生成手段

２９　分類手段

３０　マッチング・パターン生成手段

３４　繰り返し部分検出手段

３５　繰り返し情報付きマッチング・パターン生成手段

３８　スペーサ用イメージ検出手段

３９　ビュレット・イメージ検出手段

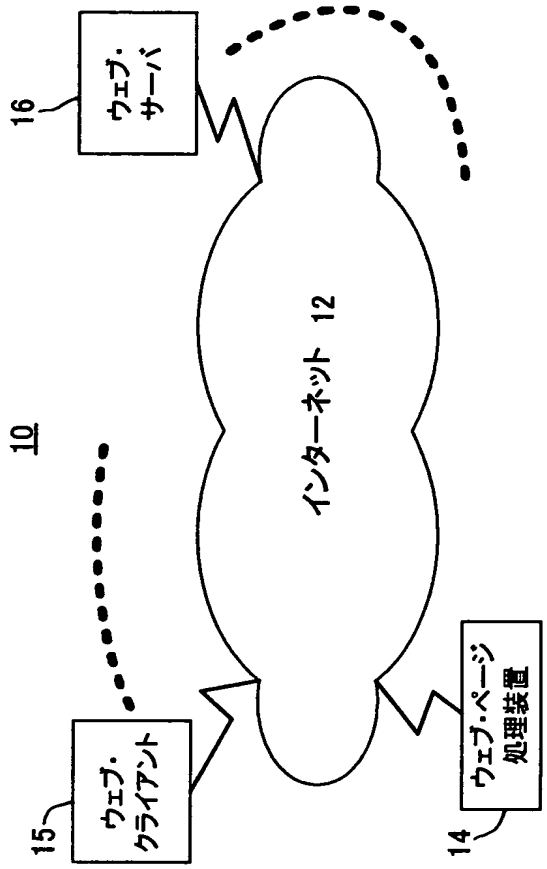
４０　第１の分類付け手段

４１　第２の分類付け手段

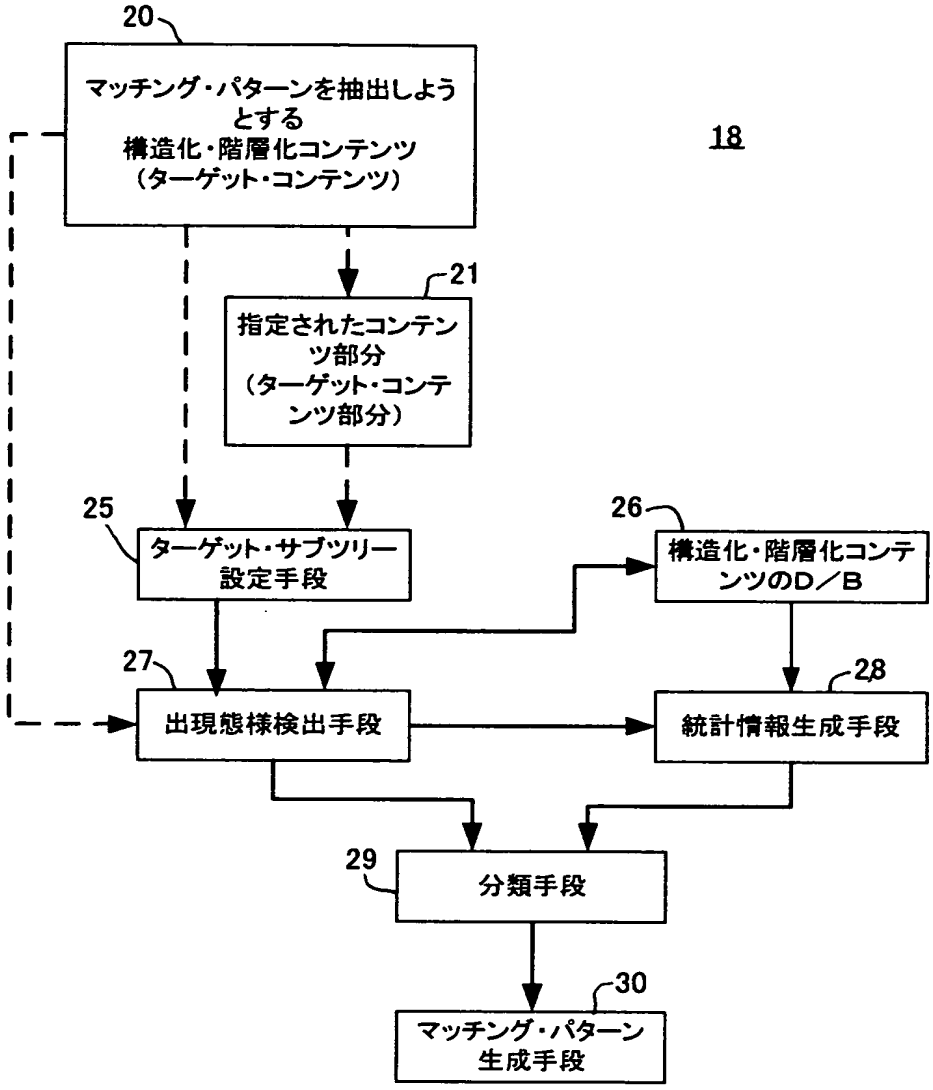
４２　分類出力手段

【書類名】　　図面

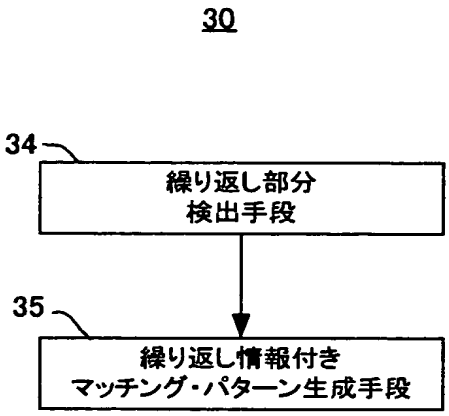
【図１】



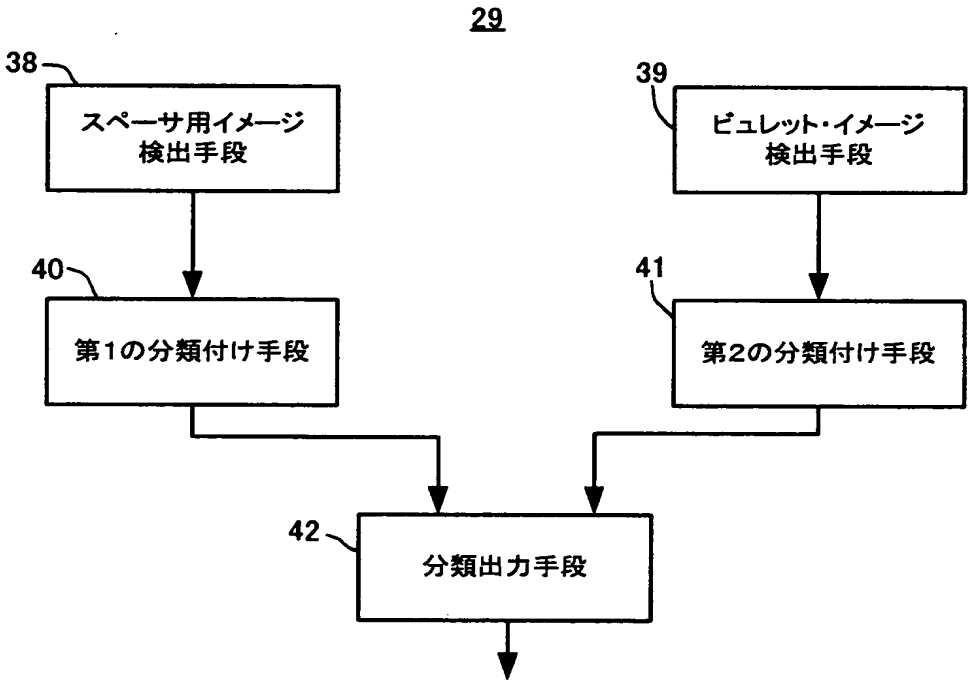
【図２】


20
マッチング・パターンを抽出しようとする構造化・階層化コンテンツ（ターゲット・コンテンツ）
18
21
指定されたコンテンツ部分（ターゲット・コンテンツ部分）
25
ターゲット・サブツリー設定手段
26
構造化・階層化コンテンツのD／B
27
出現態様検出手段
28
統計情報生成手段
29
分類手段
30
マッチング・パターン生成手段

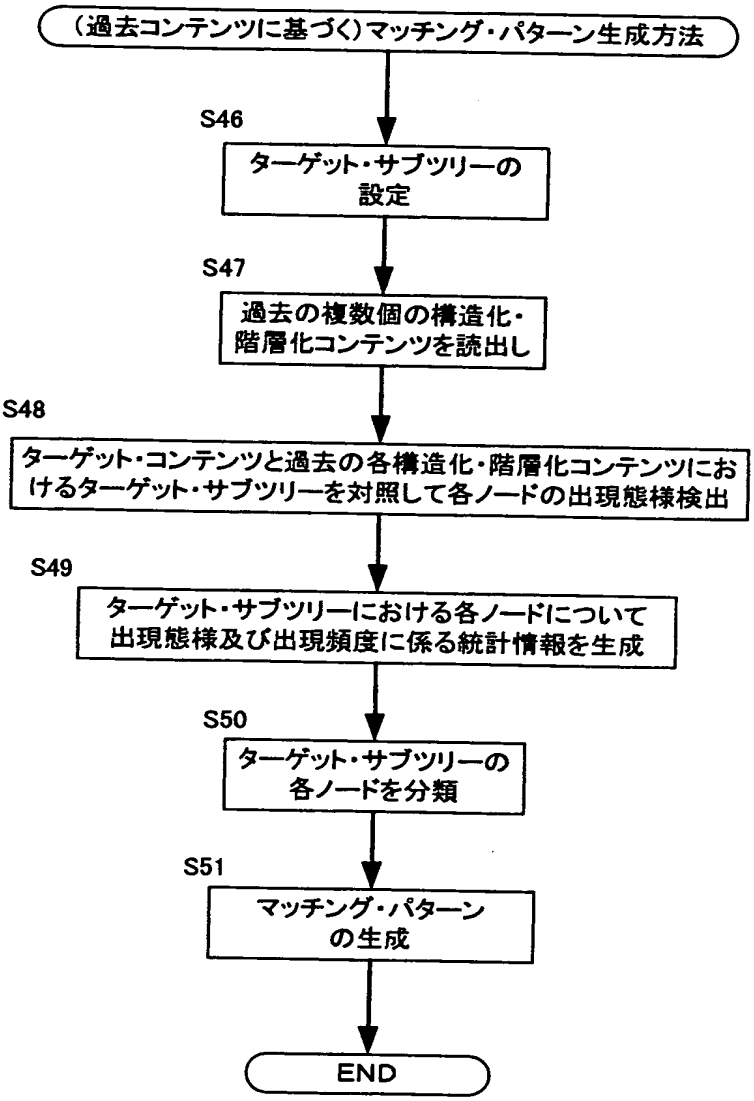
【図３】


30
34
繰り返し部分
検出手段
35
繰り返し情報付き
マッチング・パターン生成手段

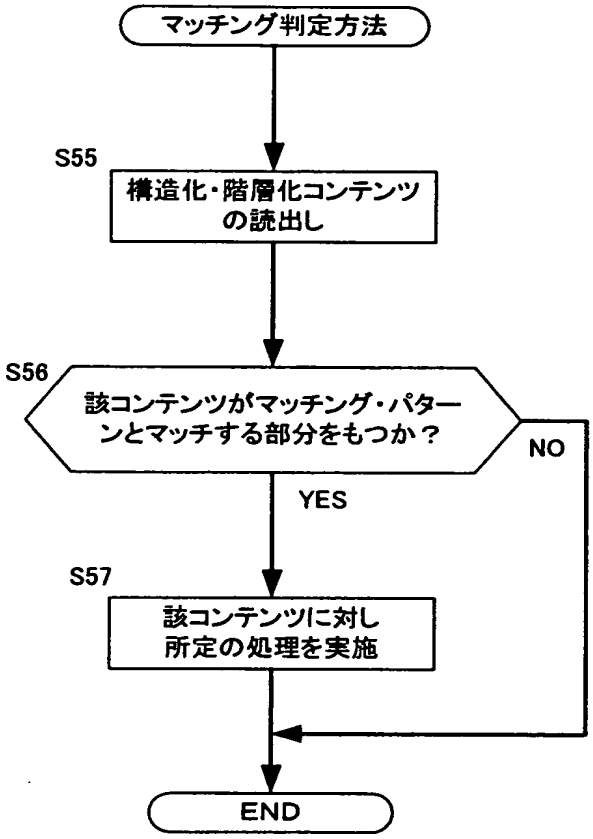
【図４】


29
38
スペーサ用イメージ
検出手段
39
ビュレット・イメージ
検出手段
40
第1の分類付け手段
41
第2の分類付け手段
42
分類出力手段

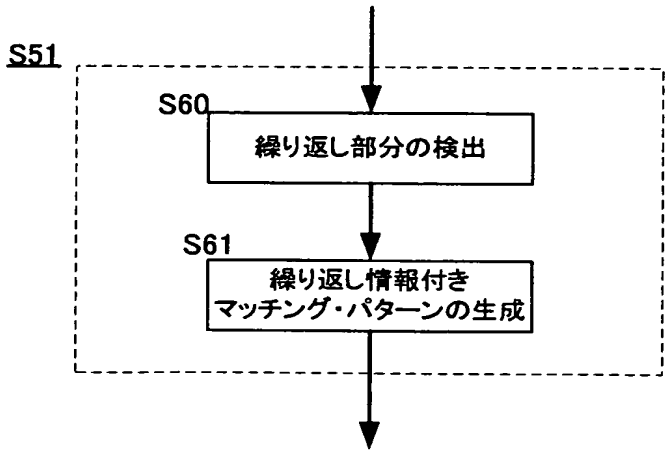
【図５】


（過去コンテンツに基づく）マッチング・パターン生成方法
S46
ターゲット・サブツリーの設定
S47
過去の複数個の構造化・階層化コンテンツを読出し
S48
ターゲット・コンテンツと過去の各構造化・階層化コンテンツにおけるターゲット・サブツリーを対照して各ノードの出現態様検出
S49
ターゲット・サブツリーにおける各ノードについて出現態様及び出現頻度に係る統計情報を生成
S50
ターゲット・サブツリーの各ノードを分類
S51
マッチング・パターンの生成
END

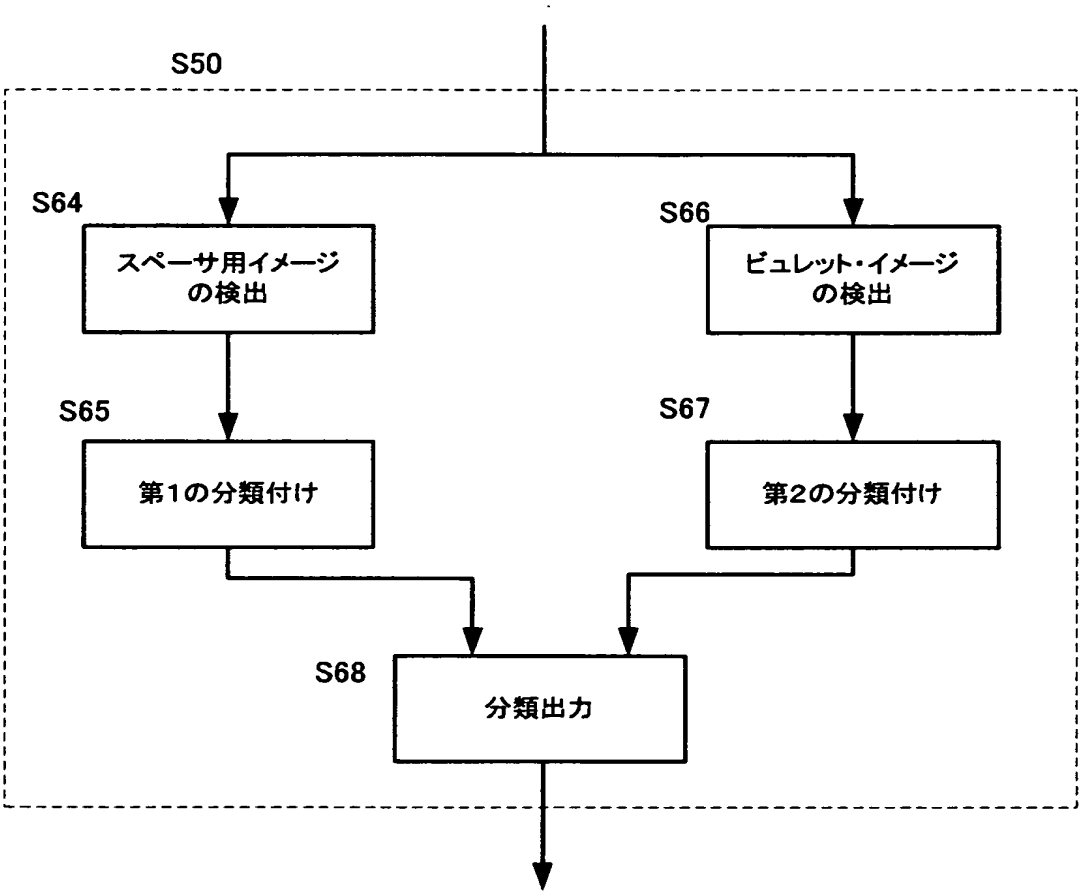
【図６】


マッチング判定方法
S55
構造化・階層化コンテンツの読出し
S56
該コンテンツがマッチング・パターンとマッチする部分をもつか？
NO
YES
S57
該コンテンツに対し所定の処理を実施
END

【図７】

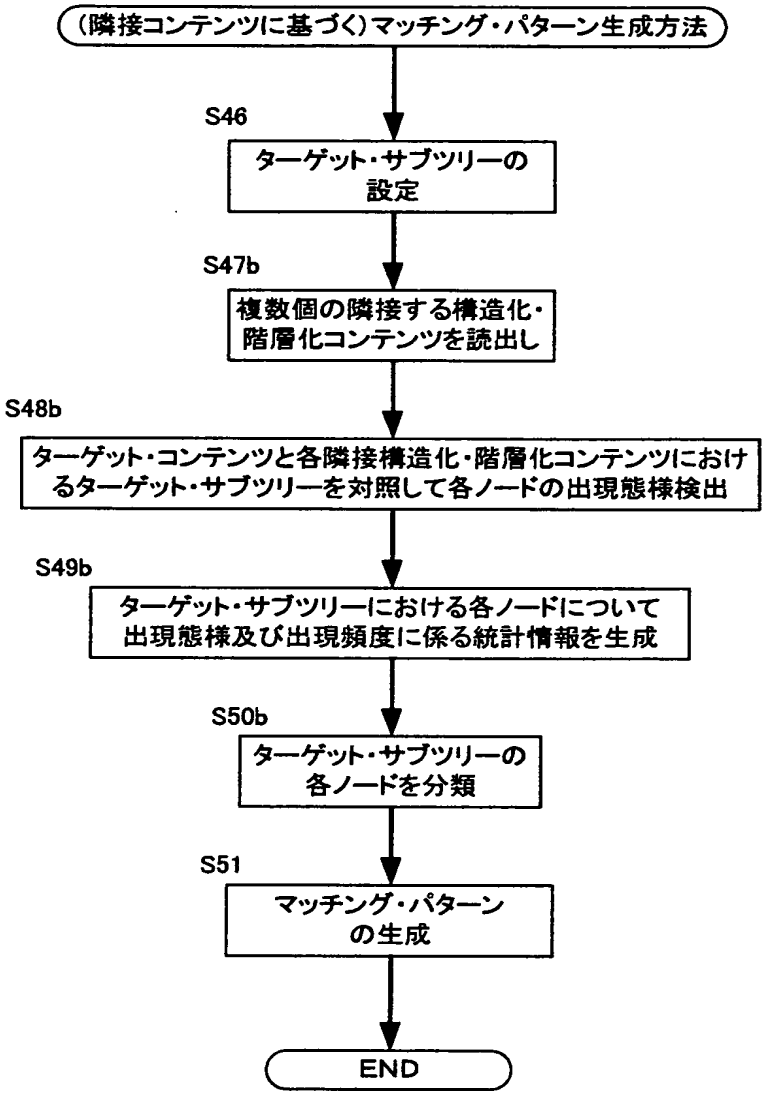

S51
S60
繰り返し部分の検出
S61
繰り返し情報付きマッチング・パターンの生成

【図８】

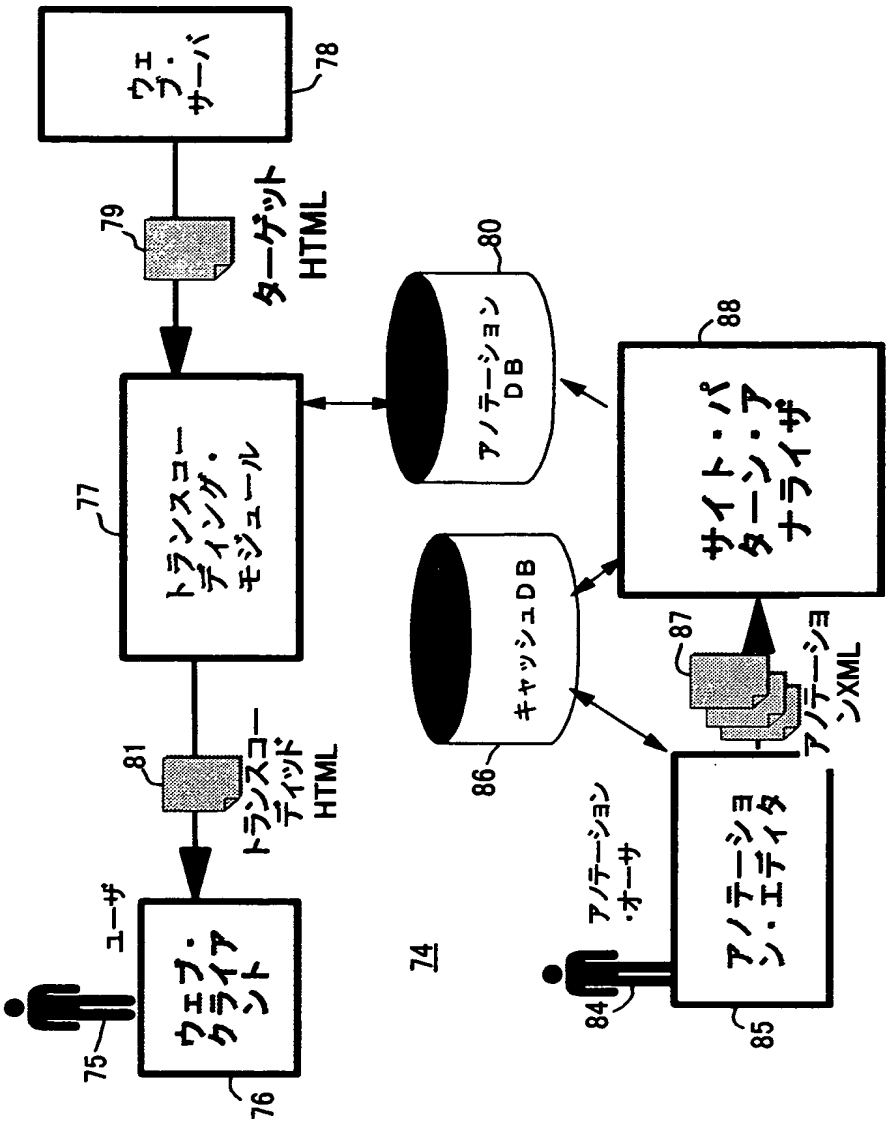


出証特２００３－３０３７００２

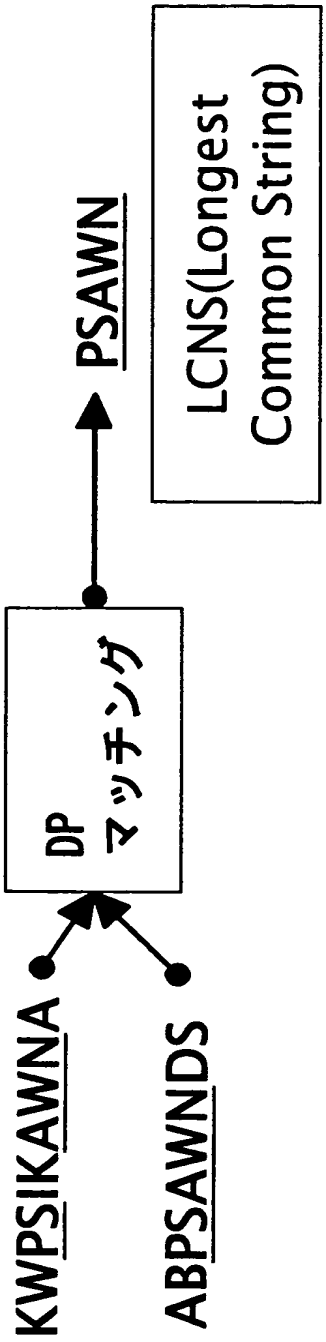
【図９】


(隣接コンテンツに基づく)マッチング・パターン生成方法
S46
ターゲット・サブツリーの設定
S47b
複数個の隣接する構造化・階層化コンテンツを読出し
S48b
ターゲット・コンテンツと各隣接構造化・階層化コンテンツにおけるターゲット・サブツリーを対照して各ノードの出現態様検出
S49b
ターゲット・サブツリーにおける各ノードについて出現態様及び出現頻度に係る統計情報を生成
S50b
ターゲット・サブツリーの各ノードを分類
S51
マッチング・パターンの生成
END

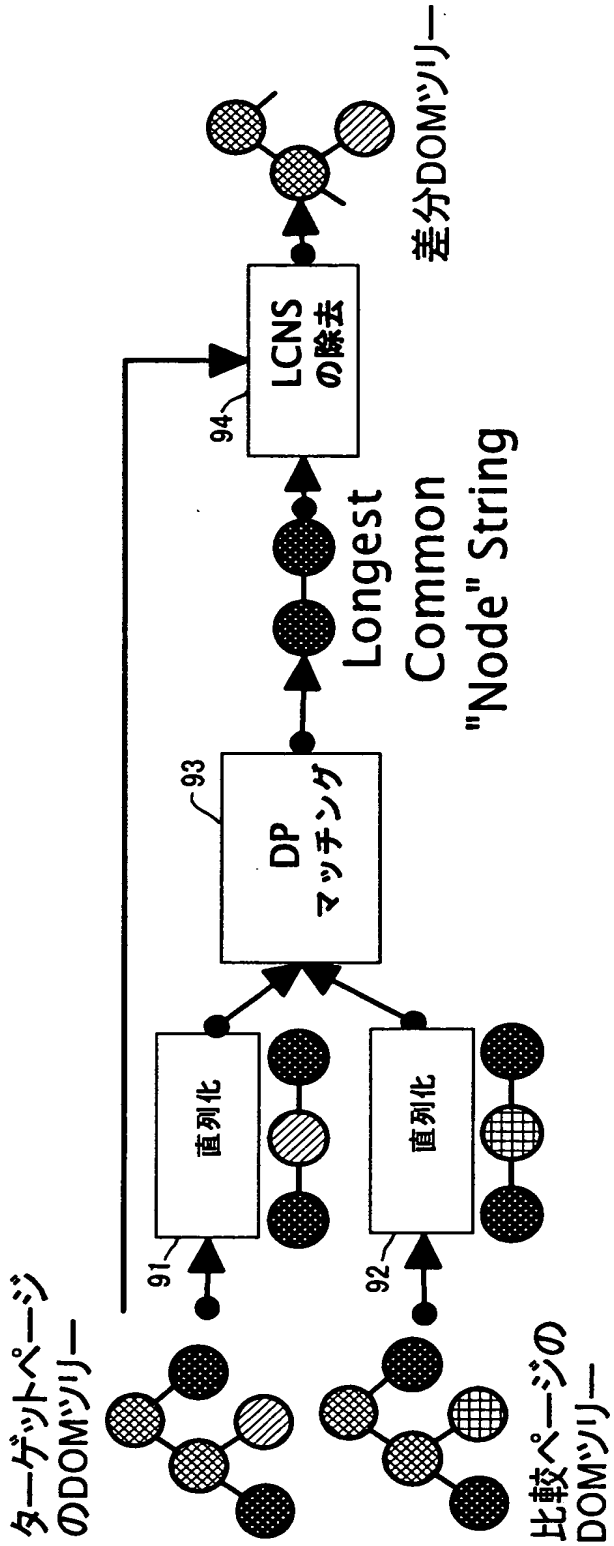
【図１０】


ウェブ・サーバ
78
79
ターゲットHTML
80
アノテーションDB
88
サイト・パターン・アナライザ
77
トランスコーディング・モジュール
キャッシュDB
87
アノテーションXML
81
トランスコーディッドHTML
86
ユーザ
ウェブ・クライアント
アノテーション・オーサ
アノテーション・エディタ
75
76
84
85
74

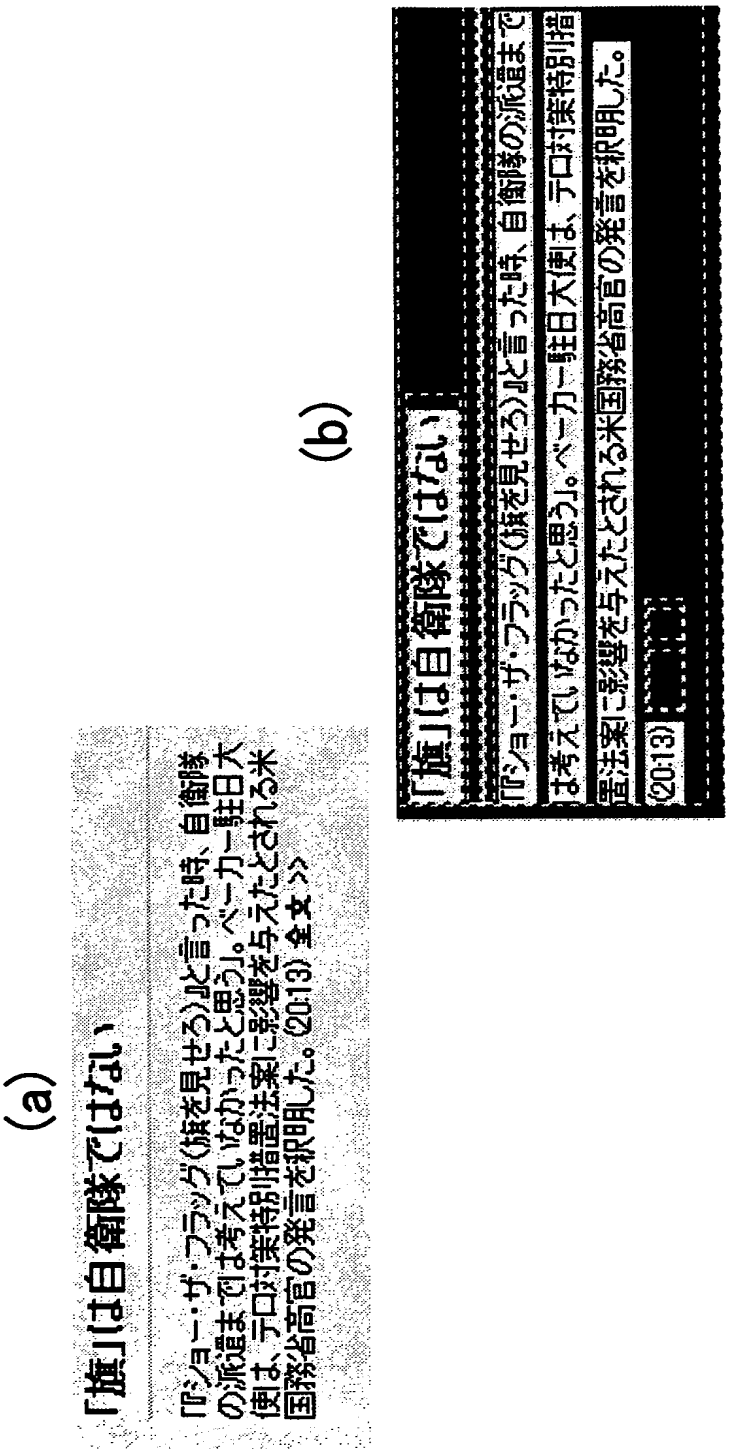
【図１１】


KWPSIKAWNA
ABPSAWNDS
DP
マッチング
PSAWN
LCNS(Longest
Common String)

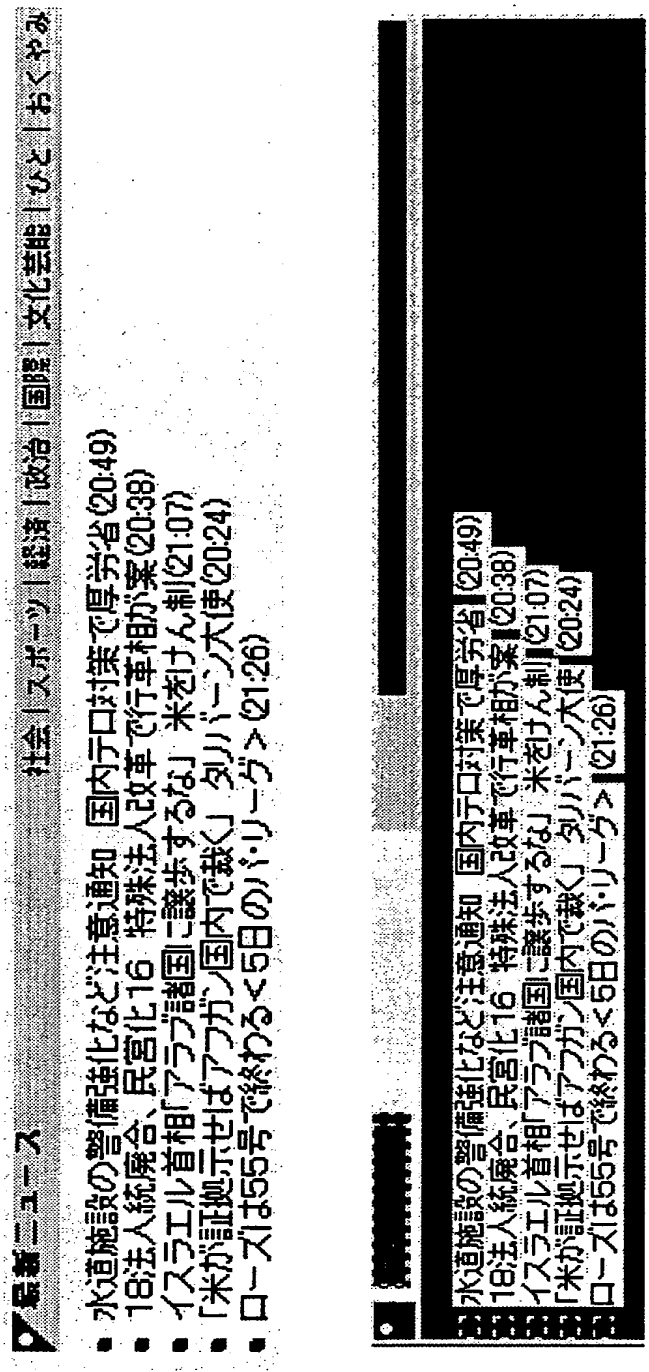
【図１２】


ターゲットページ
のDOMツリー
比較ページの
DOMツリー
91
92
直列化
直列化
93
DP
マッチング
Longest
Common
"Node" String
94
LCNS
の除去
差分DOMツリー

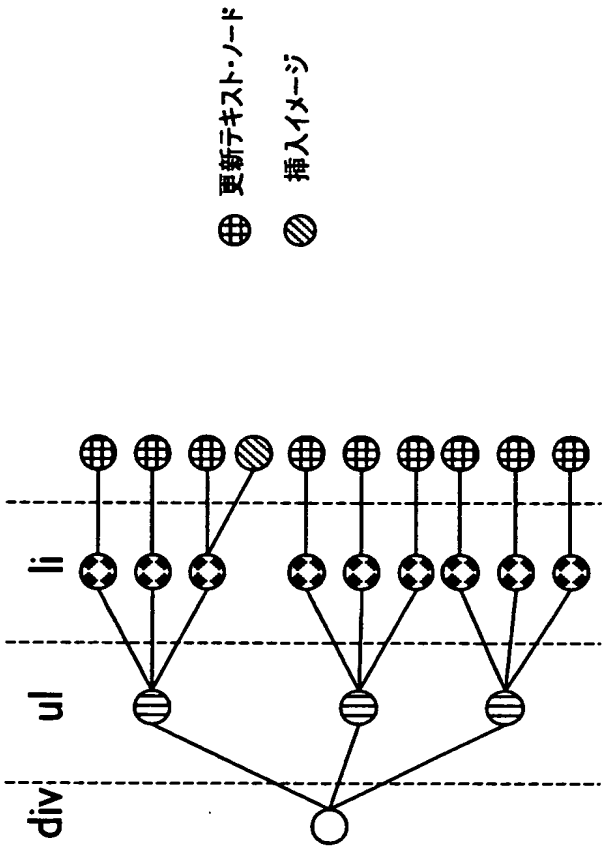
【図１３】


(a)
「旗」は自衛隊ではない
「ショー・ザ・フラッグ(旗を見せろ)」と言った時、自衛隊の派遣までは考えていなかったと思う」。ベーカー駐日大使は、テロ対策特別措置法案に影響を与えたとされる米国務省高官の発言を釈明した。(20:13) 全文 >>
(b)
「旗」は自衛隊ではない
「ショー・ザ・フラッグ(旗を見せろ)」と言った時、自衛隊の派遣までは考えていなかったと思う」。ベーカー駐日大使は、テロ対策特別措置法案に影響を与えたとされる米国務省高官の発言を釈明した。
(20:13)

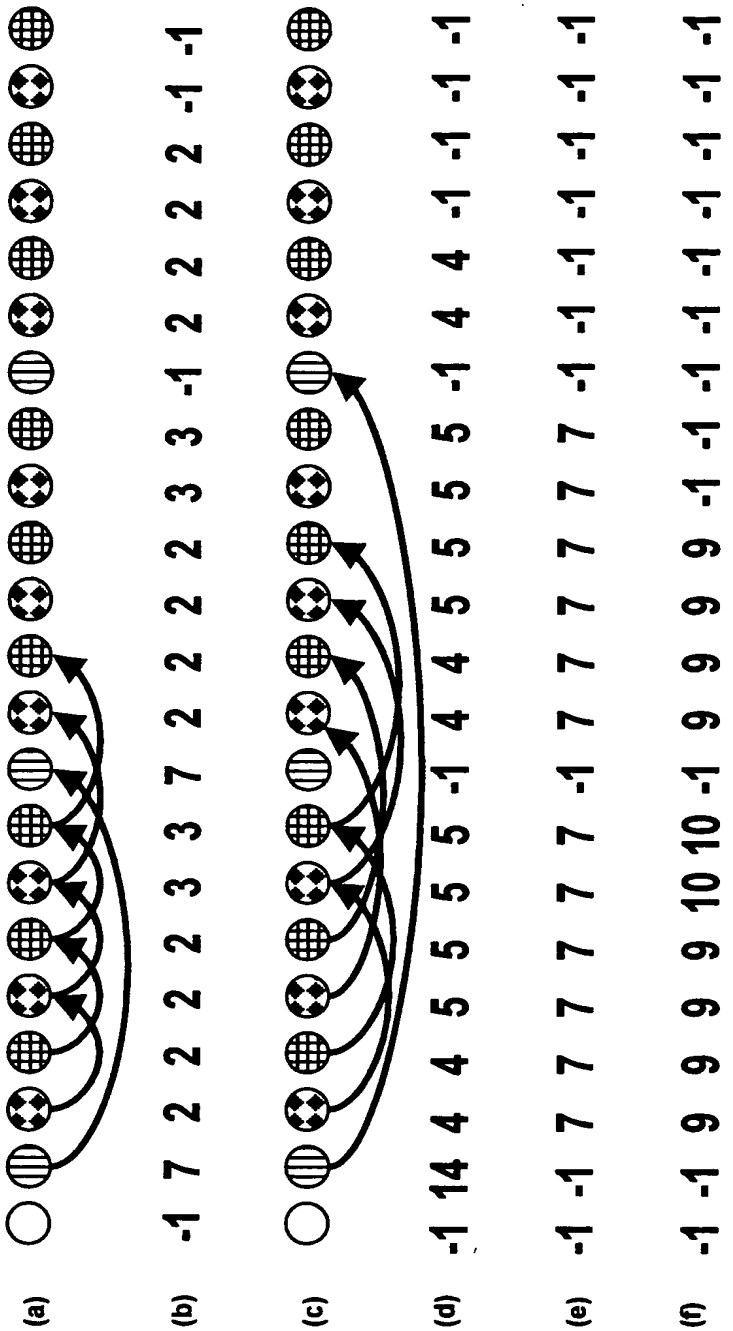
【図１４】


(a)
最新ニュース
社会 | スポーツ | 経済 | 政治 | 国際 | 文化芸能 | ひと | おくやみ
水道施設の警備強化など注意通知 国内テロ対策で厚労省(20:49)
18法人統廃合、民営化16 特殊法人改革で行革相が案(20:38)
イスラエル首相「アラブ諸国に譲歩するな」 米ぜけん制(21:07)
「米が証拠示せばアフガン国内で裁く」 タリバーン大使(20:24)
ローズは5号で終わる<5日のパ・リーグ>(21:26)
(b)
水道施設の警備強化など注意通知 国内テロ対策で厚労省(20:49)
18法人統廃合、民営化16 特殊法人改革で行革相が案(20:38)
イスラエル首相「アラブ諸国に譲歩するな」 米ぜけん制(21:07)
「米が証拠示せばアフガン国内で裁く」 タリバーン大使(20:24)
ローズは5号で終わる<5日のパ・リーグ>(21:26)

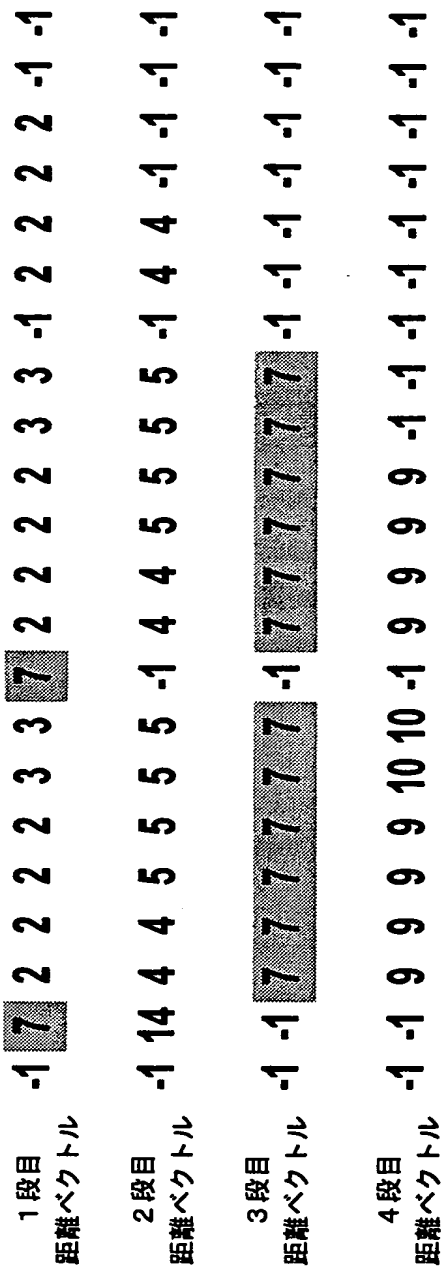
【図１５】



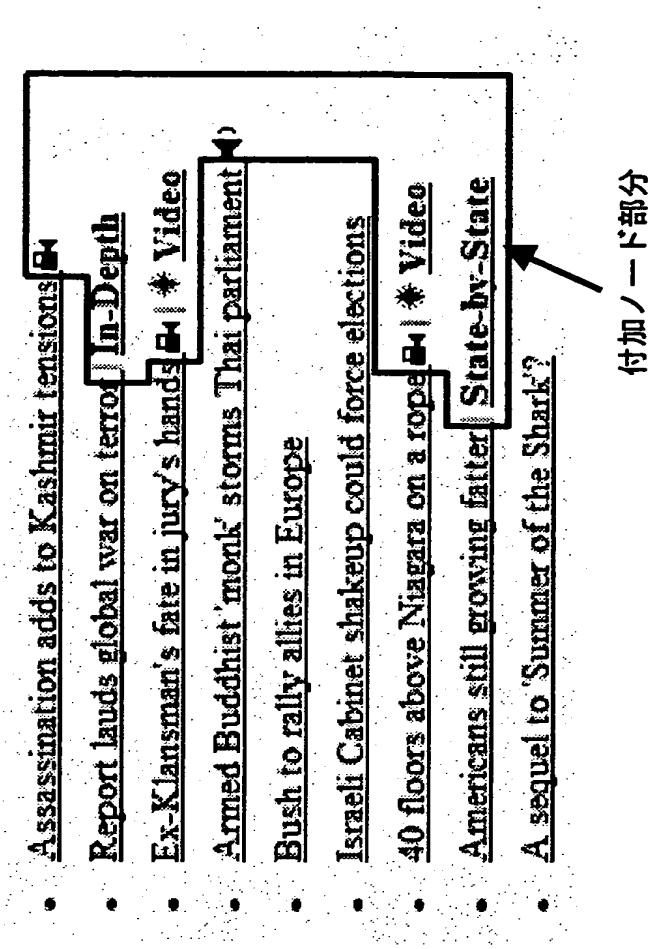
【図１６】


(a)
(b) -1 7 2 2 2 2 3 3 7 2 2 2 2 3 3 -1 2 2 2 2 -1 -1
(c)
(d) -1 14 4 4 5 5 5 5 -1 4 4 5 5 5 5 -1 4 4 -1 -1 -1 -1
(e) -1 -1 7 7 7 7 7 7 -1 7 7 7 7 7 7 -1 -1 -1 -1 -1 -1
(f) -1 -1 9 9 9 9 9 10 10 -1 9 9 9 9 9 -1 -1 -1 -1 -1 -1

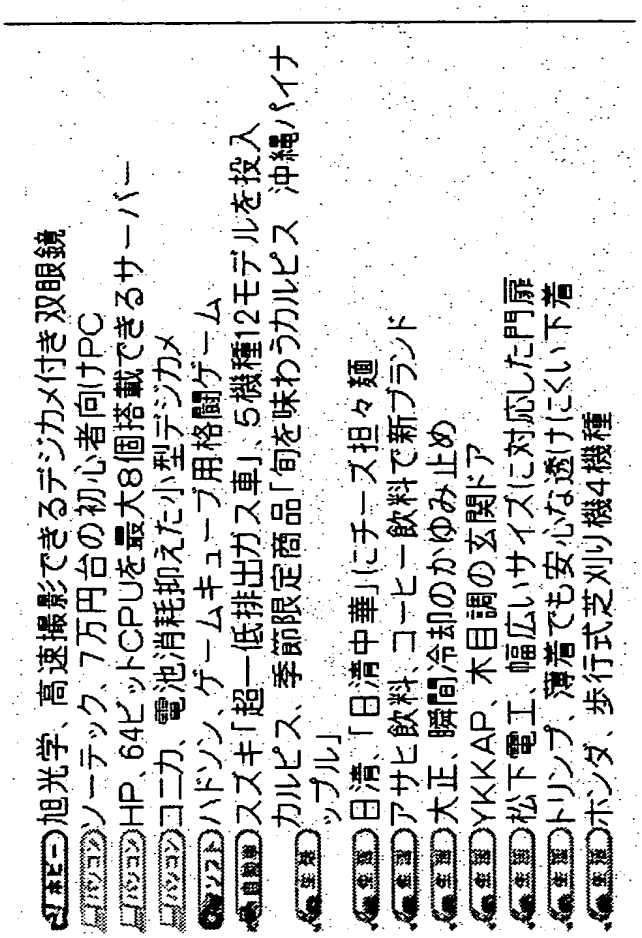
【図１７】


１段目
距離ベクトル
-1 7 2 2 2 2 3 3 7 2 2 2 2 3 3 3 -1 2 2 2 2 2 -1 -1
２段目
距離ベクトル
-1 14 4 4 5 5 5 5 -1 4 4 5 5 5 5 5 -1 4 4 -1 -1 -1 -1
３段目
距離ベクトル
-1 -1 7 7 7 7 7 7 -1 7 7 7 7 7 7 7 -1 -1 -1 -1 -1 -1
４段目
距離ベクトル
-1 -1 9 9 9 9 9 10 10 -1 9 9 9 9 9 -1 -1 -1 -1 -1 -1

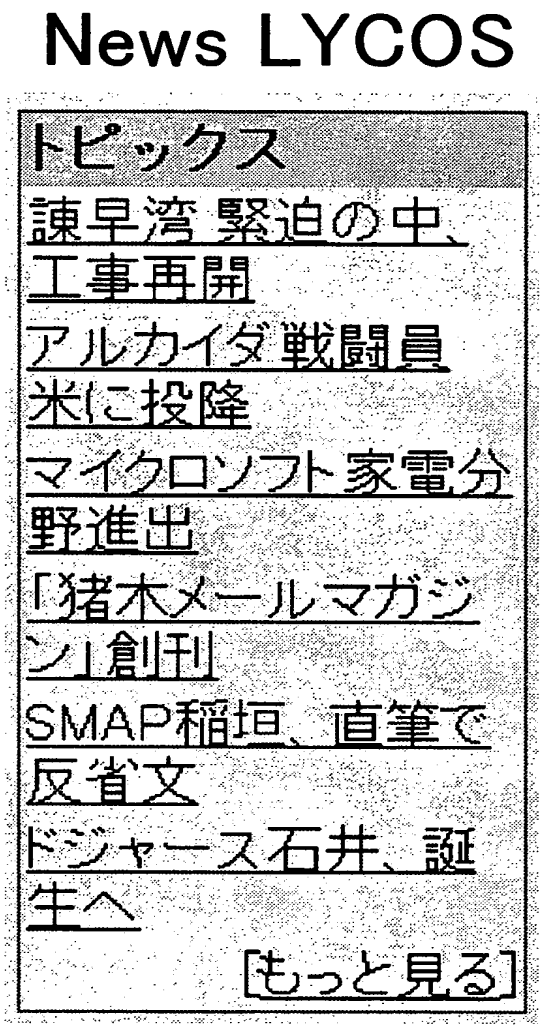
【図１８】


Assassination adds to Kashmir tensions
Report lauds global war on terror In-Depth
Ex-Klansman's fate in jury's hands Video
Armed Buddhist 'monk' storms Thai parliament
Bush to rally allies in Europe
Israeli Cabinet shakeup could force elections
40 floors above Niagara on a rope Video
Americans still growing fatter State-by-State
A sequel to 'Summer of the Shark'?
付加ノード部分

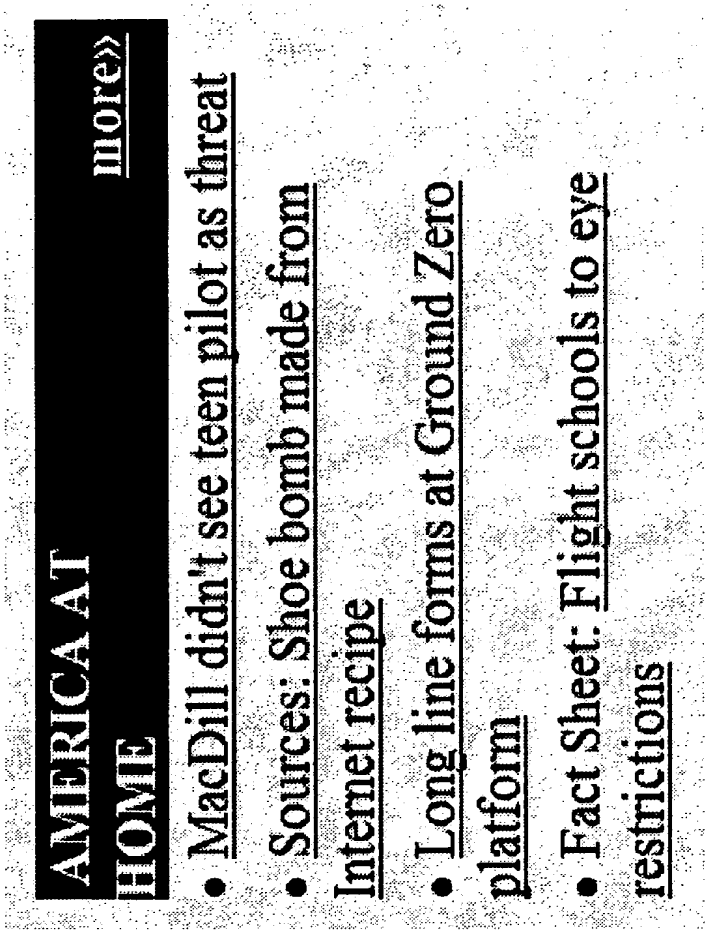
【図１９】


旭光学、高速撮影できるデジカメ付き双眼鏡
ソーテック、7万円台の初心者向けPC
HP、64ビットCPUを最大8個搭載できるサーバー
コニカ、電池消耗抑えた小型デジカメ
ハドソン、ゲームキューブ用格闘ゲーム
スズキ「超一低排出ガス車」、5機種12モデルを投入
カルピス、季節限定商品「旬を味わうカルピス 沖縄パイナップル」
日清、「日清中華」にチーズ担々麺
アサヒ飲料、コーヒー飲料で新ブランド
大正、瞬間冷却のかゆみ止め
YKKAP、木目調の玄関ドア
松下電工、幅広いサイズに対応した門扉
トリンプ、洗濯でも安心な透けにくい下着
ホンダ、歩行式芝刈り機4機種

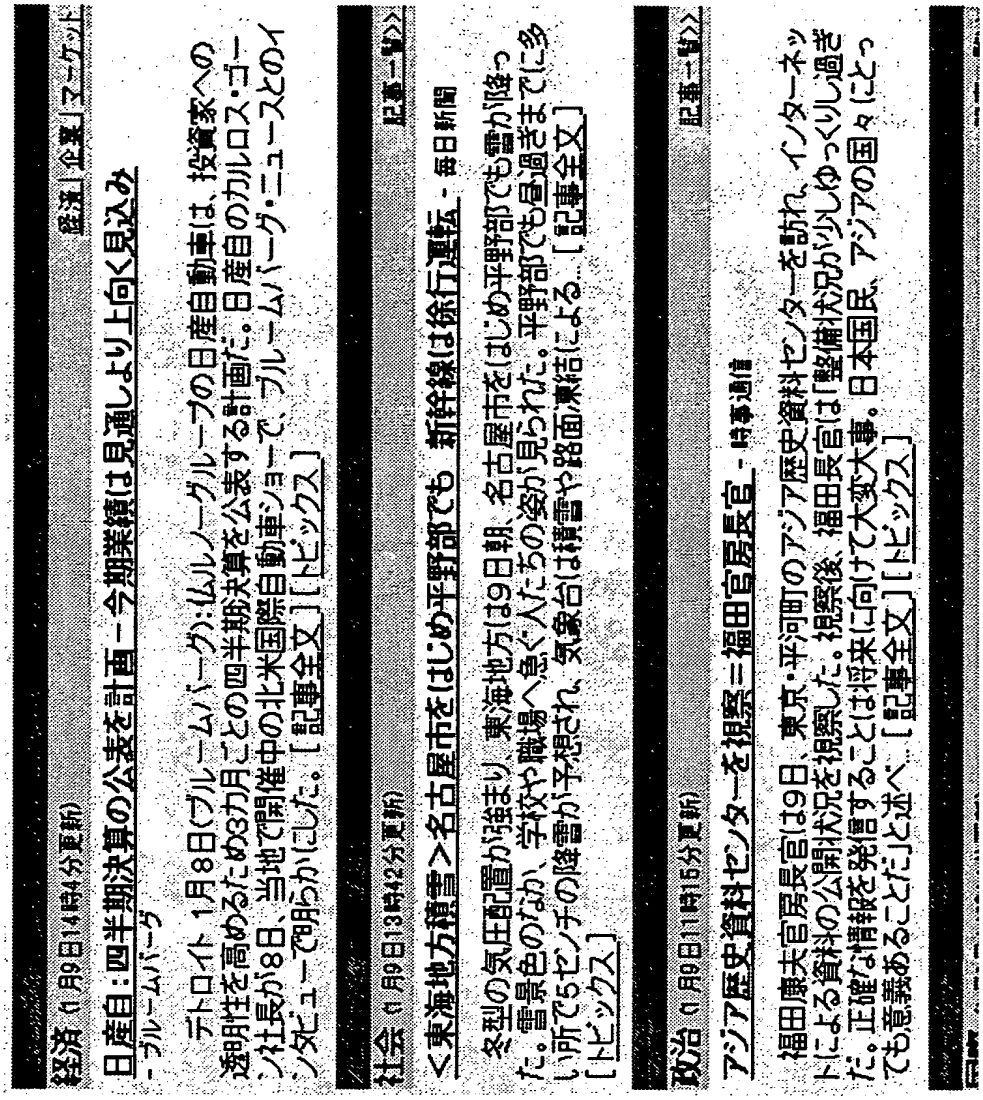
【図２０】


News LYCOS
トピックス
諫早湾 緊迫の中、工事再開
アルカイダ戦闘員 米に投降
マイクロソフト家電分野進出
「猪木メールマガジン」創刊
SMAP稲垣、直筆で反省文
ドジャース石井、誕生へ
[もっと見る]

【図２１】

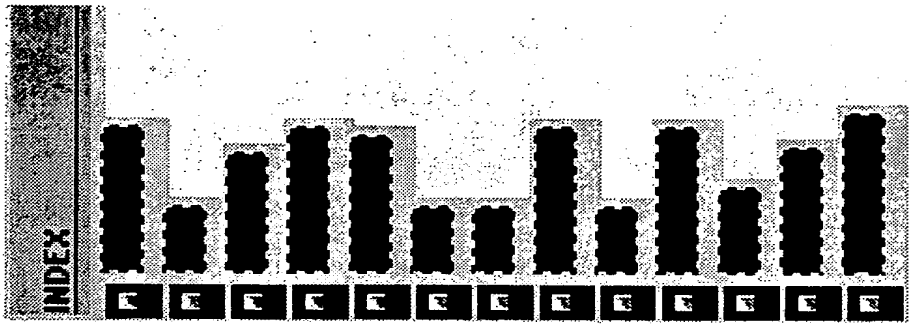

CNN.COM
AMERICA AT HOME
more»
• MacDill didn't see teen pilot as threat
• Sources: Shoe bomb made from Internet recipe
• Long line forms at Ground Zero platform
• Fact Sheet: Flight schools to eye restrictions

【図２２】

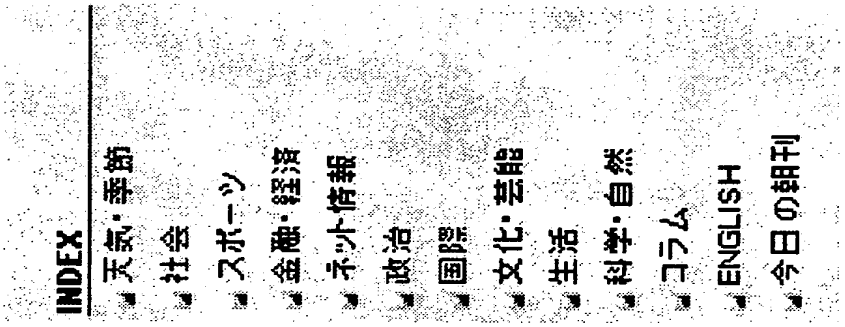

経済 (1月9日14時4分更新)
経済 | 企業 | マーケット
日産自：四半期決算の公表を計画－今期業績は見通しより上向く見込み
- ブルームバーグ
デトロイト 1月8日(ブルームバーグ):仏ルノーグループの日産自動車は、投資家への透明性を高めるため3カ月ごとの四半期決算を公表する計画だ。日産自のカルロス・ゴーン社長が8日、当地で開催中の北米国際自動車ショーで、ブルームバーグ・ニュースとのインタビューで明らかにした。[記事全文] [トピックス]
社会 (1月9日13時42分更新)
記事一覧>>
<東海地方積雪>名古屋市をはじめ平野部でも 新幹線は徐行運転 - 毎日新聞
冬型の気圧配置が強まり、東海地方は9日朝、名古屋市をはじめ平野部でも雪が降った。雪景色のなか、学校や職場へ急ぐ人たちの姿が見られた。平野部でも昼過ぎまでに多い所で5センチの降雪が予想され、気象台は積雪や路面凍結による... [記事全文] [トピックス]
政治 (1月9日11時15分更新)
記事一覧>>
アジア歴史資料センターを視察＝福田官房長官 - 時事通信
福田康夫官房長官は9日、東京・平河町のアジア歴史資料センターを訪れ、インターネットによる資料の公開状況を視察した。視察後、福田長官は「整備が少しゆっくりし過ぎだ。正確な情報を発信することは将来に向けて大変大事。日本国民、アジアの国々にとっても意義あることだ」と述べ... [記事全文] [トピックス]

【図２３】

(b)


INDEX

(a)


INDEX
天気・季節
社会
スポーツ
金融・経済
ネット情報
政治
国際
文化・芸能
生活
科学・自然
コラム
ENGLISH
今日の朝刊

【図２４】

ローズとの勝負避けたダイエーに抗議相次ぐ

9月30日のプロ野球ダイエー近鉄最終戦（福岡ドーム）で、ダイエーが本塁打の日本新記録を狙うローズ外野手との勝負を避けた問題が、波紋を広げている。日本プロ野球機構の川島広守コミッショナーは1日、「そのような野球がファンに支持されるとは到底思えない」と不快感を示す談話を発表。ダイエー球団にも抗議、質問の電話やメールが計数十件届いた。

ローズ外野手は30日、ダイエー・王貞治監督の持つシーズン最多本塁打55本に並んでダイエー戦を迎えた。しかしダイエー側は、若菜バッテリーコーチの指示で、ローズに対しストライクをほとんど投げなかった。

試合中、ダイエーファンからも「なぜ堂々と勝負しないのか」と怒りの声が浴びせられた。

これを受けて同コミッショナーは1日、談話を発表。「（記録達成の機会を故意に奪うことは）フェアプレーを至上の価値とする野球の本質から外れている。そうして守られた記録は、その記録ばかりか記録を達成した選手の人格をも汚すことになる」と訴えた。

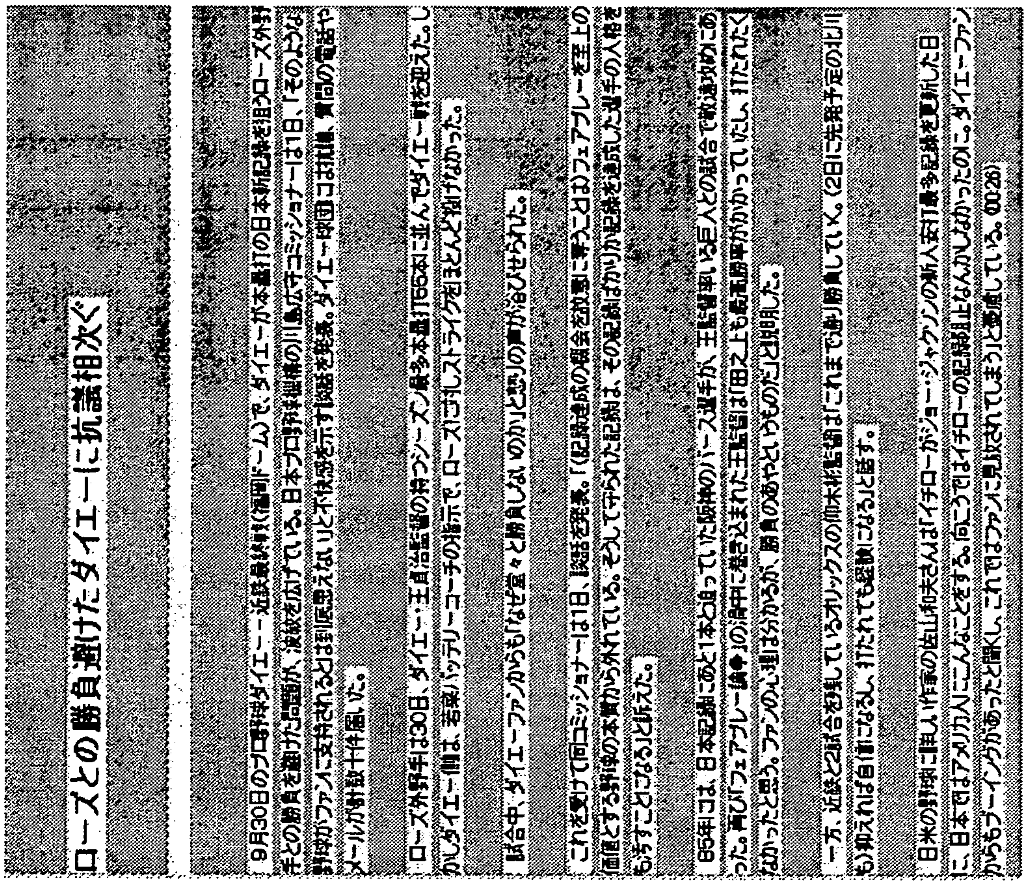
85年には、日本記録にあと1本と迫っていた阪神のバース選手が、王監督率いる巨人との試合で敬遠攻めにあった。再び「フェアプレー論争」の渦中に巻き込まれた王監督は「田之上も最高勝率がかかっていたし、打たれたくなかったと思う。ファンの心理は分かるが、勝負のあやというものだ」と説明した。

一方、近鉄と2試合を残しているオリックスの仰木彬監督は「これまで通り勝負していく。（2日に先発予定の北川も）抑えれば自信になるし、打たれても経験になる」と話す。

日米の野球に詳しい作家の佐山和夫さんは「イチローがジョー・ジャクソンの新人安打最多記録を更新した日に、日本ではアメリカ人にこんなことをする。向こうではイチローの記録阻止なんかしなかったのに。ダイエーファンからもブーイングがあったと聞くし、これではファンに見放されてしまう」と憂慮している。（00-26）

www.asahi.com/sports/update/1001/009.html

【図２５】



ローズとの勝負避けたダイエーに抗議相次ぐ

9月30日のプロ野球ダイエー－近鉄最終戦(福岡ドーム)で、ダイエーが本塁打の日本新記録を狙うローズ外野手との勝負を避けた問題が、波紋を広げている。日本プロ野球組織の川島広守コミッショナーは1日、「そのような野球がファンに支持されるとは到底思えない」と不快感を示す談話を発表。ダイエー球団には抗議、質問の電話やメールが計数十件届いた。

ローズ外野手は30日、ダイエー・王貞治監督の持つシーズン最多本塁打55本に並んでダイエー戦を迎えた。しかしダイエー側は、若菜バッテリーコーチの指示で、ローズにはストライクをほとんど投げなかった。

試合中、ダイエーファンからも「なぜ堂々と勝負しないのか」と怒りの声が浴びせられた。

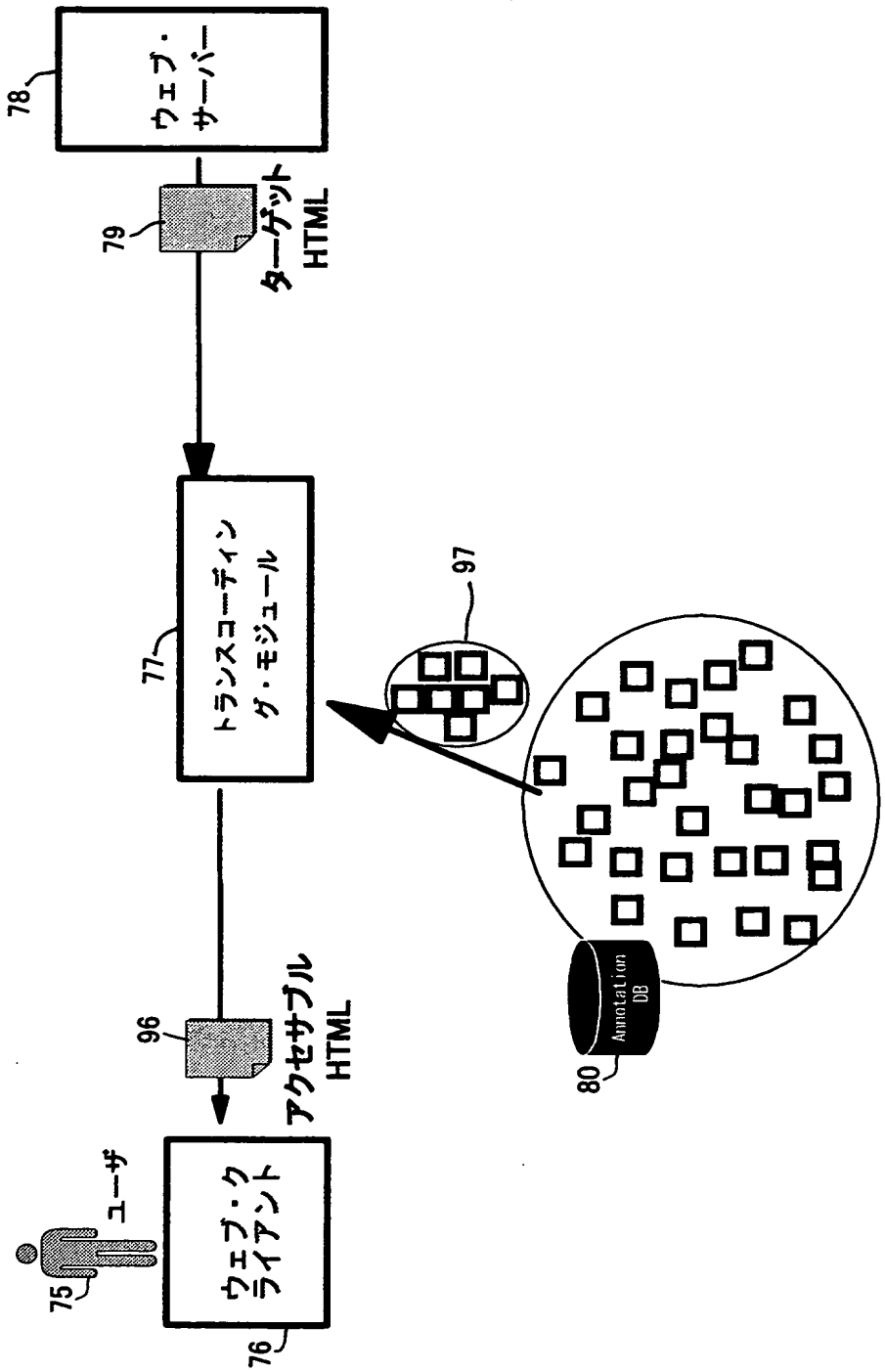
これを受けて同コミッショナーは1日、談話を発表。「(記録達成の機会を故意に奪うことは)フェアプレーを至上の価値とする野球の本質から外れている。そうして守られた記録は、その記録ばかりか記録を達成した選手の人格をも汚すことになる」と訴えた。

85年には、日本記録にあと1本と迫っていた阪神のバース選手が、王監督率いる巨人との試合で敬遠攻めにあった。再び「フェアプレー論争」の渦中に巻き込まれた王監督は「田之上も最高勝率がかかっていたし、打たれたくなかったと思う。ファンの心理は分かるが、勝負のあやというものだ」と説明した。

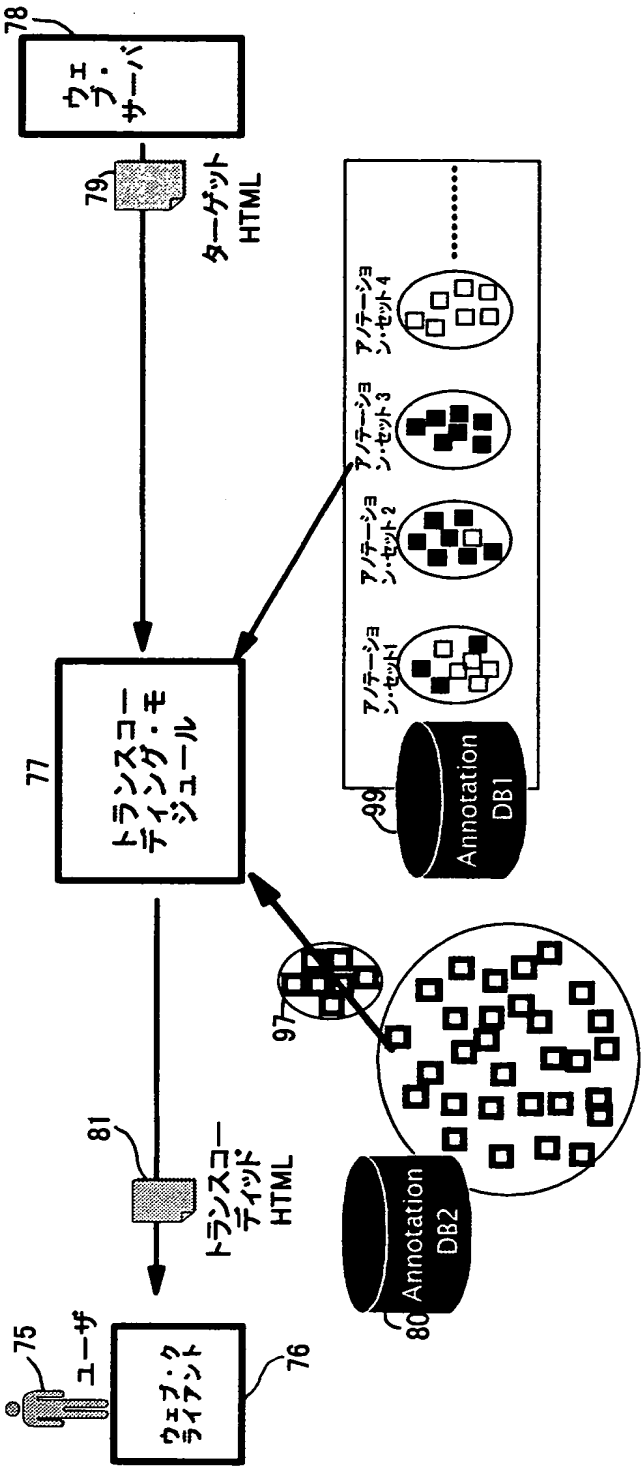
一方、近鉄と2試合を残しているオリックスの仰木彬監督は「これまで通り勝負していく。(2日に先発予定の北川も)抑えれば自信になるし、打たれても経験になる」と話す。

日米の野球に詳しい作家の佐山和夫さんは「イチローがジョー・ジャクソンの新人安打最多記録を更新した日に、日本ではアメリカ人にこんなことをする。向こうではイチローの記録阻止なんかしなかったのに。ダイエーファンからもブーイングがあったと聞くし、これではファンに見放されてしまう」と憂慮している。(00:26)

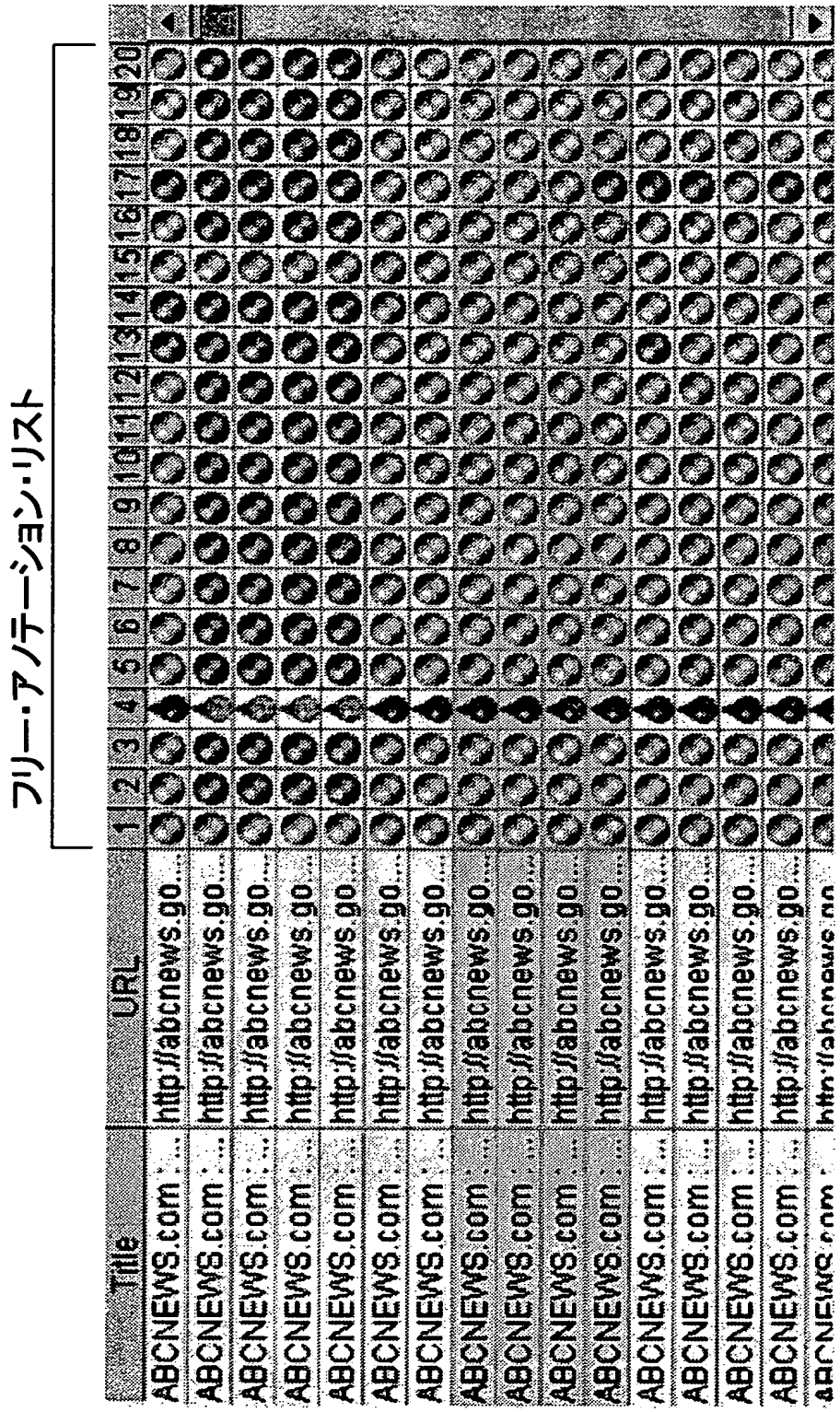
【図２６】


78
ウェブ・サーバー
79
ターゲットHTML
77
トランスコーディング・モジュール
97
Annotation DB
80
96
アクセサブルHTML
ユーザ
75
ウェブ・クライアント
76

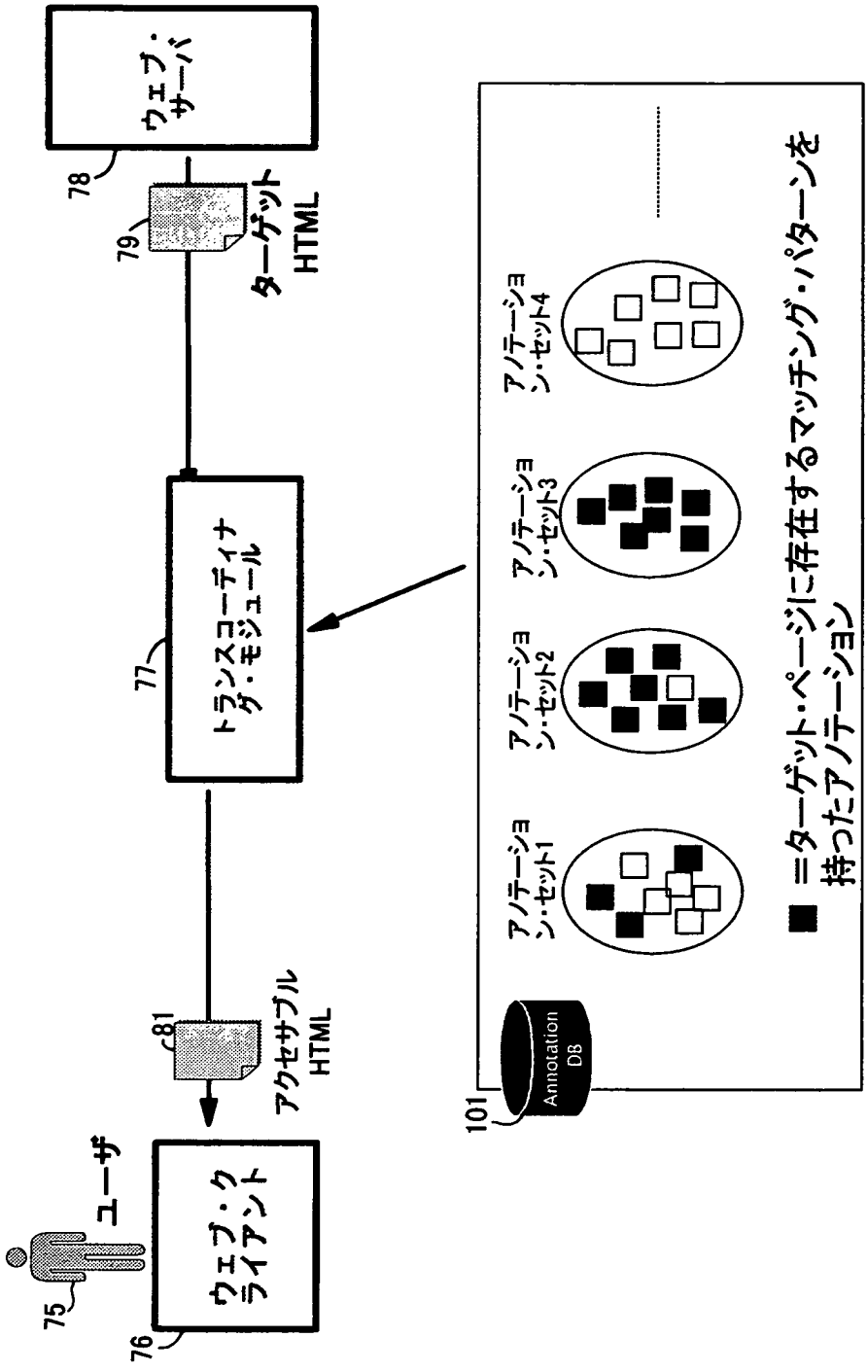
【図２７】


78
ウェブ・サーバ
79
ターゲットHTML
アノテーション・セット4
アノテーション・セット3
アノテーション・セット2
アノテーション・セット1
99
Annotation DB1
77
トランスコーディング・モジュール
97
81
トランスコーディッドHTML
Annotation DB2
80
75
ユーザ
ウェブ・クライアント
76

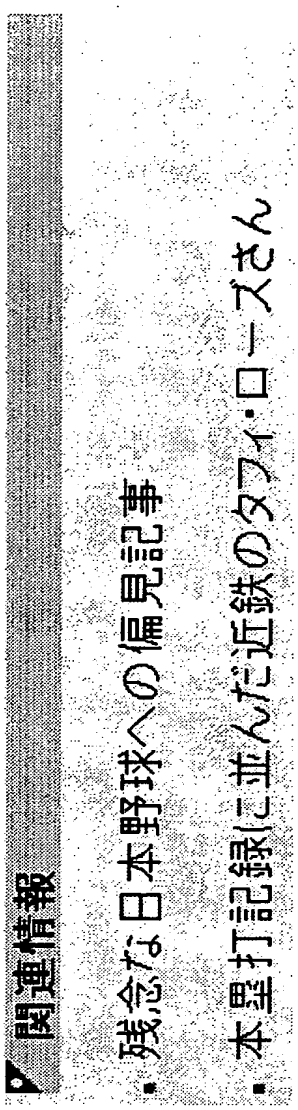
【図２８】


フリー・アノテーション・リスト
Title
URL
ABCNEWS.com :...
http://abcnews.go....
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
URLリスト

【図２９】


ウェブ・サーバ
78
79
ターゲット HTML
トランスコーディング・モジュール
77
81
アクセサブル HTML
ウェブ・クライアント
76
ユーザ
75
101
Annotation DB
アノテーション・セット1
アノテーション・セット2
アノテーション・セット3
アノテーション・セット4
■ ターゲット・ページに存在するマッチング・パターンを持つアノテーション

【図３０】


関連情報
・残念な日本野球への偏見記事
・本塁打記録に並んだ近鉄のタフィ・ローズさん

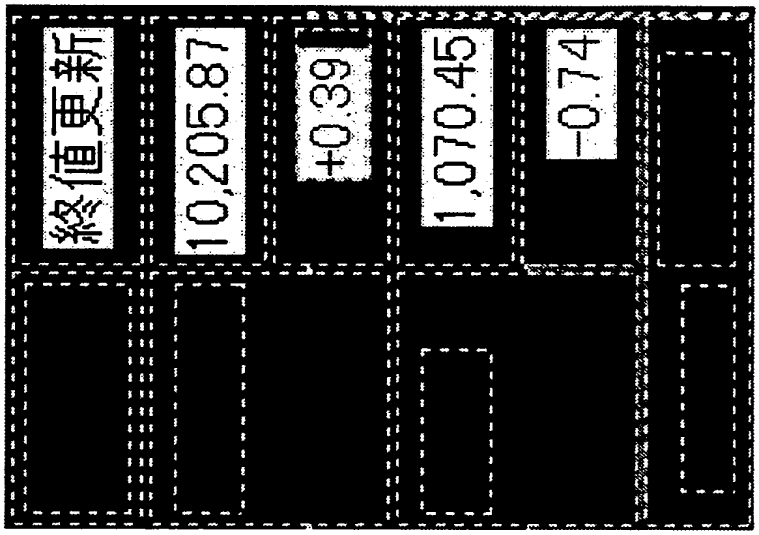
(a)


関連情報
・残念な日本野球への偏見記事
・本塁打記録に並んだ近鉄のタフィ・ローズさん

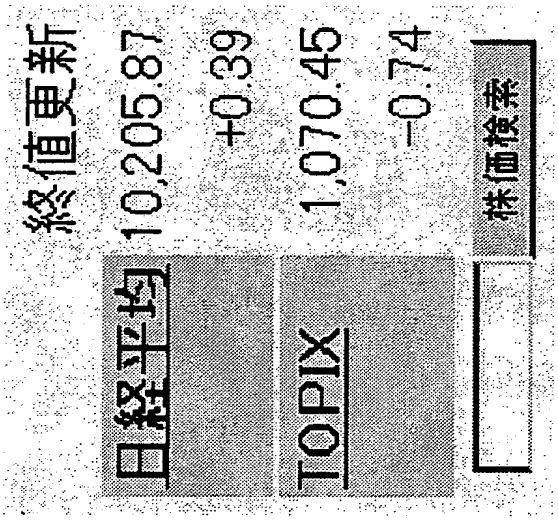
(b)

【図３１】

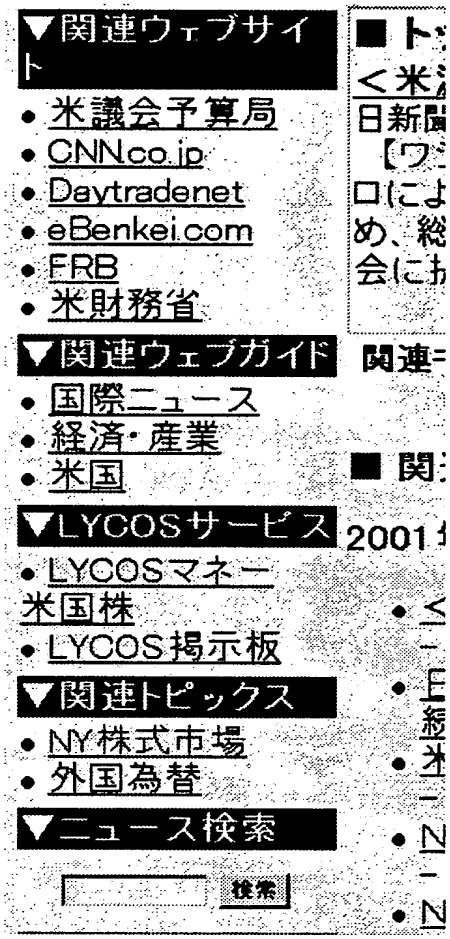
(b)


終値更新
10,205.87
+0.39
1,070.45
−0.74

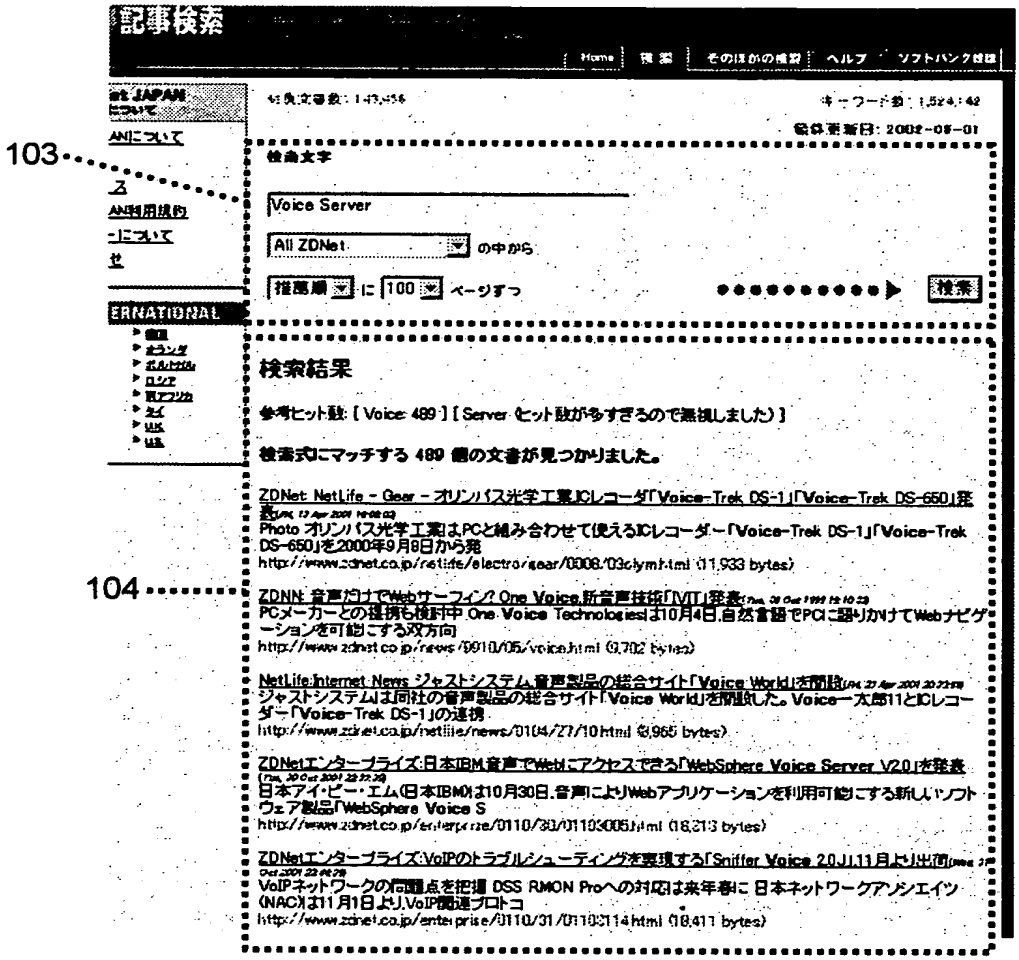
(a)


終値更新
日経平均 10,205.87 +0.39
TOPIX 1,070.45 −0.74
株価検索

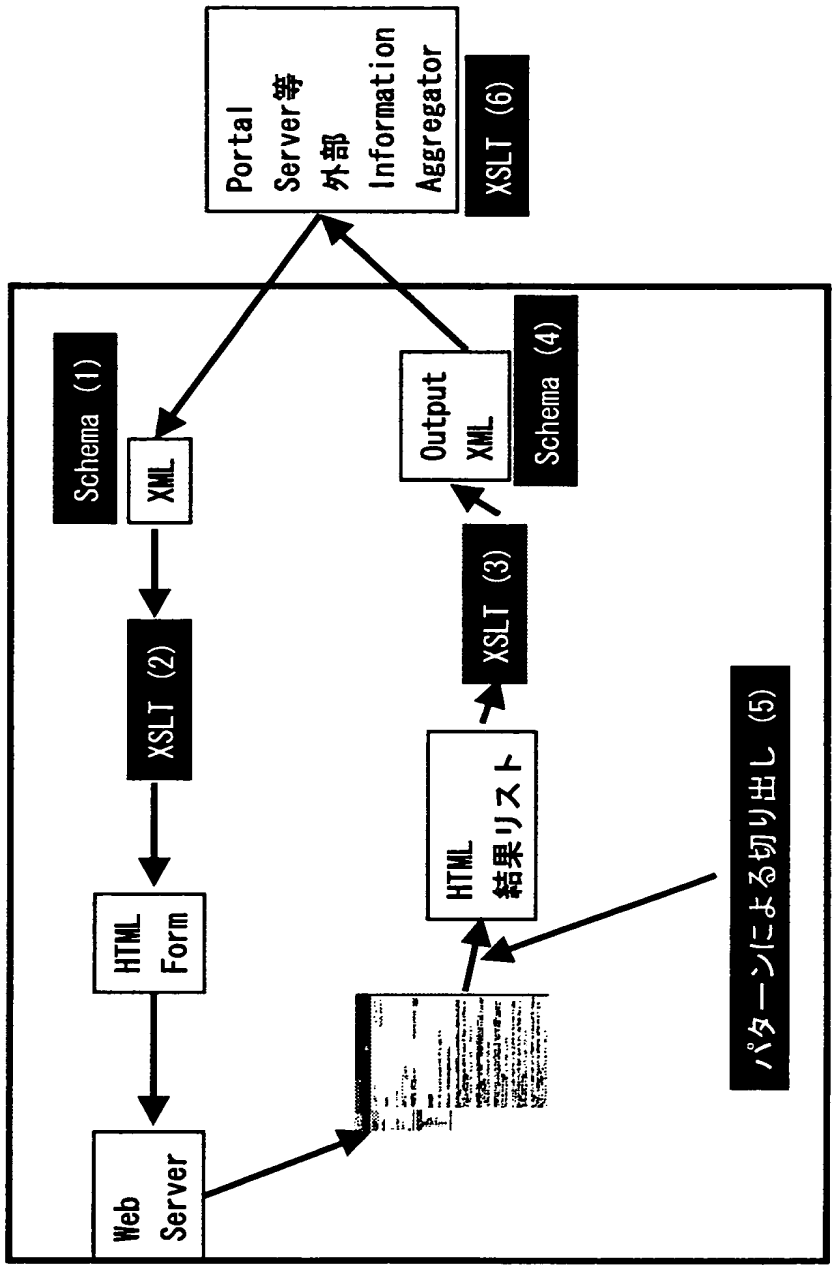
【図３２】


▼関連ウェブサイト
• 米議会予算局
• CNN.co.jp
• Daytradenet
• eBenkei.com
• FRB
• 米財務省
▼関連ウェブガイド
• 国際ニュース
• 経済・産業
• 米国
▼LYCOSサービス
• LYCOSマネー
米国株
• LYCOS掲示板
▼関連トピックス
• NY株式市場
• 外国為替
▼ニュース検索
検索

【図３３】


記事検索
Home
検索
そのほかの検索
ヘルプ
ソフトバンク検索
103
Voice Server
All ZDNet
の中から
に
100
ページずつ
検索
検索結果
参考ヒット数: [Voice: 489] [Server: ヒット数が多すぎるので無視しました)]
検索式にマッチする 489 個の文書が見つかりました。
104
ZDNet: NetLife - Gear - オリンパス光学工業ICレコーダ「Voice-Trek DS-1」「Voice-Trek DS-650」発表
Photo オリンパス光学工業はPCと組み合わせて使えるICレコーダー「Voice-Trek DS-1」「Voice-Trek DS-650」を2000年9月8日から発
http://www.zdnet.co.jp/netlife/electro/gear/0008/03olymp.html (11,933 bytes)
ZDNN 音声だけでWebサーフィン? One Voice,新音声技術「IVIT」発表
PCメーカーとの提携も検討中 One Voice Technologiesは10月4日,自然言語でPCに語りかけてWebナビゲーションを可能にする双方向
http://www.zdnet.co.jp/news/9910/05/voice.html (9,702 bytes)
NetLife:Internet News ジャストシステム,音声製品の総合サイト「Voice World」を開設
ジャストシステムは同社の音声製品の総合サイト「Voice World」を開設した。Voice一太郎11とICレコーダー「Voice-Trek DS-1」の連携
http://www.zdnet.co.jp/netlife/news/0104/27/10.html (8,965 bytes)
ZDNetエンタープライズ:日本IBM,音声でWebにアクセスできる「WebSphere Voice Server V2.0」を発表
日本アイ・ビー・エム(日本IBM)は10月30日,音声によりWebアプリケーションを利用可能にする新しいソフトウェア製品「WebSphere Voice S
http://www.zdnet.co.jp/enterprise/0110/30/0110300S.html (18,213 bytes)
ZDNetエンタープライズ:VoIPのトラブルシューティングを実現する「Sniffer Voice 2.0」,11月より出荷
VoIPネットワークの問題点を把握 DSS RMON Proへの対応は来年春に 日本ネットワークアソシエイツ(NAC)は11月1日より,VoIP関連プロトコ
http://www.zdnet.co.jp/enterprise/0110/31/0110310114.html (18,411 bytes)

【図３４】


Portal Server等 外部 Information Aggregator
XSLT (6)
Schema (1)
XML
XSLT (2)
HTML Form
Web Server
Output XML
Schema (4)
XSLT (3)
HTML 結果リスト
パターンによる切り出し (5)

【書類名】　　要約書

【要約】

【課題】　XPathを用いずに構造化・階層化コンテンツの一部切り出し等の処理を効率的に行う。

【解決手段】　ターゲット・サブツリー設定手段２５は、コンテンツ部分２１に係るターゲット・サブツリーを設定する。出現態様検出手段２７は、コンテンツ２０に係るターゲット・サブツリーと過去の各構造化・階層化コンテンツに係るツリーとを対照してターゲット・サブツリーの各ノードの出現態様を検出する。統計情報生成手段２８はターゲット・サブツリーにおける各ノードについての出現態様の出現頻度に係る統計情報を生成する。分類手段２９は、出現態様検出結果及び統計情報に基づいてターゲット・サブツリーの各ノードを分類する。マッチング・パターン生成手段３０は該分類に基づいてターゲット・コンテンツ部分に係るマッチング・パターンを生成する。該マッチング・パターンを使って、構造化・階層化コンテンツを識別する。

【選択図】　　図２

認定・付加情報

特許出願の番号	特願２００２－３１２３３１
受付番号	５０２０１６２０９７０
書類名	特許願
担当官	末武　実　　１９１２
作成日	平成１４年１２月　６日

＜認定情報・付加情報＞

【提出日】　平成14年10月28日

【特許出願人】

【識別番号】　390009531

【住所又は居所】　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】　100086243

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　市位　嘉宏

【代理人】

【識別番号】　100108501

【住所又は居所】　神奈川県大和市下鶴間１６２３番１４　日本アイ・ビー・エム株式会社　知的所有権

【氏名又は名称】　上野　剛史

【復代理人】　申請人

【識別番号】　100085408

【住所又は居所】　東京都中央区日本橋２丁目１番１号　櫻正宗ビル９階

【氏名又は名称】　山崎　隆

次頁無

出　願　人　履　歴　情　報

識別番号　　〔３９０００９５３１〕

１．変更年月日　　２００２年　６月　３日
［変更理由］　住所変更
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレーション